

Doctoral Thesis

Software-Driven E-commerce: Enhancing Usability, User  
Experience, Accessibility, and Traffic Optimization through  
Machine Learning, Natural Language Processing, Large  
Language Models, and Search Engine Optimization  
Techniques



**Konstantinos I. Roumeliotis**

University of Peloponnese

Department of Informatics and Telecommunications

Faculty of Economics and Technology

Tripoli, May 2024

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ηλεκτρονικό εμπόριο με γνώμονα το λογισμικό:  
Βελτιστοποίηση Ευχρηστίας, Εμπειρίας Χρήστη,  
Προσβασιμότητας και Επισκεψιμότητας βάσει  
Μηχανικής Μάθησης, Επεξεργασίας Φυσικής Γλώσσας,  
Μεγάλων Γλωσσικών Μοντέλων και τεχνικών  
Βελτιστοποίησης Μηχανών Αναζήτησης



**Κωνσταντίνος Ι. Ρουμελιώτης**  
Πανεπιστήμιο Πελοποννήσου  
Τμήμα Πληροφορικής και Τηλεπικοινωνιών  
Σχολή Οικονομίας και Τεχνολογίας

Τρίπολη, Μάιος 2024

## **Internal Advisory Committee**

- Tselikas Nikolaos, Professor, Department of Informatics and Telecommunications, University of the Peloponnese
- Vassilakis Costas, Professor, Department of Informatics and Telecommunications, University of the Peloponnese
- Tryfonopoulos Christos, Professor, Department of Informatics and Telecommunications, University of the Peloponnese

## **External Examination Committee**

- Lepouras George, Professor, Department of Informatics and Telecommunications, University of the Peloponnese
- Kokkinos Panagiotis, Associate Professor, Department of Digital Systems, University of the Peloponnese
- Prezerakos George, Professor, Department of Informatics and Computer Engineering, University of West Attica
- Nasiopoulos Dimitrios, Associate Professor, Department of Agribusiness and Supply Chain Management, Agricultural University of Athens

© Konstantinos I. Roumeliotis, 2024

*Series of dissertations submitted to the  
Faculty of Economics and Technology, University of Peloponnese*

*All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.*



## **Abstract**

This thesis examines the techniques and technologies that can lead to a more optimized, more accessible, and more sustainable WEB and E-commerce. The development of software tools in both PHP and Python programming languages is undertaken, leveraging advanced Large Language Models (LLMs) and Natural Language Processing (NLP) to automate E-commerce processes previously deemed inaccessible. To confirm and enhance the results of the research, data analysis tools, predictive modeling, and Machine Learning (ML) algorithms were utilized. Throughout this research, all the aforementioned technologies are harmoniously combined, leading to an E-commerce that will not only survive but thrive in the future's intense competition.

This research is divided into four distinct chapters, each specialized in one of the aforementioned technologies. The investigation begins with LLMs, exploring ways to integrate them into E-commerce and how their advanced Artificial Intelligence (AI) and NLP capabilities can automate E-commerce processes. Subsequently, an exploration of Web Accessibility is undertaken, highlighting its status as an emerging domain for future consideration within the virtual landscape. This occurs notwithstanding the escalating demand for an enhanced level of accessibility on the web. Progressive Web Apps (PWA), a new web technology relying on cutting-edge technologies to transform web pages, particularly E-commerce, into a multi-device tool, increasing accessibility and usability, are then explored. Subsequently, an evaluation of PWAs regarding accessibility and their ability to deliver on promises is conducted. Subsequent to the ongoing research, each aspect of Search Engine Optimization (SEO) is systematically examined, exploring opportunities within diverse domains under E-commerce umbrella, including retail E-commerce and service-oriented sectors such as booking platforms. This involves identifying SEO techniques and technologies that exhibit heightened efficacy in securing elevated rankings on search engines and fostering increased organic traffic. Another technology, Accelerated Mobile Pages (AMP), and how it contributes to the growth of an E-commerce's visibility is finally explored.

During the exploration of LLMs, the initial focus was on investigating NLP and its capacity for understanding human language, the Generative Pre-trained Transformer (GPT) architecture, and its innovative use of transformers and self-attention mechanisms to process input sequences. This encompassed the pre-training phase involving billion parameters utilized in the training of GPT models, as well as their capability to undergo fine-tuning for domain-specific tasks. Subsequently, specific attention was given to GPT-3.5, GPT-4, and LLaMA-2 models. Their integration into E-commerce was scrutinized, and potential enhancements, such as automation, were explored to significantly improve both functional and customer-centered aspects of online commerce. It's important to note that this exploration went beyond LLMs, delving into renowned NLP models such as BERT and RoBERTa. Additionally, unsupervised and supervised learning algorithms like k-means clustering, content-based filtering (CBF), hierarchical clustering, as well as logistic regression and neural network algorithms were examined. To achieve the research objectives, Chrome Apps and flask-based APIs were developed using Python and JavaScript. The aforementioned models underwent fine-tuning through few-shot learning tailored for specific domains, providing valuable insights into the integration of LLMs and NLP within the realm of E-commerce. The focus extended to hot topics, including sentiment analysis, recommender systems, sustainable purchasing decisions, and churn modeling. This comprehensive examination aimed to uncover the practical applications and benefits of leveraging advanced language models for optimizing the E-commerce landscape.

In the examination of Web Accessibility, through a critical review of Web Content Accessibility Guidelines (WCAG), each design principle, guideline, check point and success criterion were identified and presented in detail. Limitations in WCAG were identified in terms of both Accessibility Guidelines and efforts made by web developers and machine auditing tools. As the investigation progressed, a shift in focus occurred towards an emerging web app technology known as PWAs. PWAs rely on a blend of cutting-edge technologies such as service workers, app shell architecture, web app manifest, and caching storage API for offline functionality. In a parallel critical evaluation, the advantages and

limitations of adopting this innovative technology were explored. A sample of PWA websites and E-commerce platforms was selected, and a comparative analysis against their non-PWA counterparts was manually conducted employing popular accessibility evaluation and auditing tools. Valuable insights regarding the accessibility of PWAs compared to non-PWA websites were derived following sample collection and a descriptive analysis utilizing Jupyter and Python. This study underscores the imperative role of accessibility in the web landscape.

During the systematic literature review (SLR) in SEO, the necessity of SEO techniques and technologies for the correct presence of E-commerces on the World Wide Web and for higher rankings in the Search Engine Result Pages (SERPs) is established. The research is not limited to SEO Techniques; instead, the use of existing web technologies to boost SEO is suggested. For the purpose of the research, prototype tools in PHP and Python with crawler-like features were developed. These tools are designed to analyze targeted web pages, extracting valuable insights into the SEO techniques utilized on those pages. To simulate real-world conditions, these tools were connected to APIs, allowing them to retrieve additional metrics such as the number of backlinks, Domain Authority (DA), keywords, search ranking positions, responsive design, and website speed. Moreover, a pre-trained model was engineered through the integration of ML and predictive algorithms into the software. This model facilitated the precise prediction of the requisite number of backlinks and DA essential for a website to attain first-page ranking in search results, all while considering the intricacies of its competitive landscape. These particular tools have been released as open-source on GitHub, providing non-specialized SEO teams and business owners with the opportunity to implement SEO strategies and observe results through the software's recommendations. To confirm the tools' effectiveness and obtain more concrete insights into which SEO techniques are more efficient, case studies were conducted on specific domains within live E-commerce platforms.

Following the research, the rapid growth of the mobile phone industry was observed, leading to the exploration of alternative methods for developing web apps that are more mobile-friendly. One technology that captured the interest of webmasters was AMP. By utilizing existing technologies and imposing certain restrictions, AMP aimed to create pre-loaded web pages with advantages in terms of load time. These pages not only offered a reduction in page size but also improved rankings in search results, resulting in increased organic traffic due to their speed and user-friendliness. To ascertain whether the promised speed and ranking benefits are indeed delivered by AMP, a comparative analysis on the pages of a live E-commerce site was conducted in comparison to the corresponding pages of the same E-commerce site created using AMP technology.

***Keywords:** Accelerated Mobile Pages, Churn Modeling, Customer Satisfaction, E-commerce, Few-shot Learning, Fine-tuning, Large Language Models, Machine Learning, Natural Language Processing, Organic Traffic, Pre-trained models, Progressive Web Apps, Recommender Systems, Search Engine Optimization, Unsupervised Learning, Web Accessibility*

## Περίληψη

Η παρούσα διατριβή εξετάζει τις τεχνικές και τεχνολογίες που μπορούν να οδηγήσουν σε ένα πιο βελτιστοποιημένο, πιο προσβάσιμο και πιο βιώσιμο παγκόσμιο ιστό και ηλεκτρονικό εμπόριο. Για τους σκοπούς της έρευνας, αναπτύχθηκαν εργαλεία λογισμικού τόσο σε γλώσσα προγραμματισμού PHP όσο και σε Python, ενώ χρησιμοποιήθηκαν προηγμένα Μεγάλα Γλωσσικά Μοντέλα (LLMs) και μοντέλα Επεξεργασίας Φυσικής Γλώσσας (NLP) για την αυτοματοποίηση διαδικασιών ηλεκτρονικού εμπορίου που προηγουμένως θεωρούνταν δυσπρόσιτες. Για να επιβεβαιωθούν και να ενισχυθούν τα αποτελέσματα της έρευνας, χρησιμοποιήθηκαν εργαλεία ανάλυσης δεδομένων, προγνωστική μοντελοποίηση και αλγόριθμοι μηχανικής μάθησης. Καθ' όλη τη διάρκεια αυτής της έρευνας, όλες οι προαναφερόμενες τεχνολογίες συνδυάζονται αρμονικά, οδηγώντας σε ένα E-commerce που όχι μόνο θα επιβιώσει αλλά θα ακμάσει στον έντονο ανταγωνισμό του μέλλοντος.

Η παρούσα έρευνα διαιρείται σε τέσσερα διακριτά κεφάλαια, το καθένα εξειδικευμένο σε μία από τις προαναφερθείσες τεχνολογίες. Η έρευνα ξεκινά με τα Μεγάλα Γλωσσικά Μοντέλα (LLMs), εξερευνώντας τρόπους ενσωμάτωσής τους στο ηλεκτρονικό εμπόριο και πώς οι προηγμένες ικανότητες Τεχνητής Νοημοσύνης (TN) και Επεξεργασίας Φυσικής Γλώσσας μπορούν να αυτοματοποιήσουν τις διαδικασίες του ηλεκτρονικού εμπορίου. Στη συνέχεια, γίνεται μια έρευνα της Προσβασιμότητας στον Ιστό, επισημαίνοντας τη θέση της ως αναδυόμενου πεδίου για το μέλλον. Έπειτα, εξετάζονται οι Προοδευτικές Εφαρμογές Ιστού (PWA), μια νέα τεχνολογία ιστού που βασίζεται σε προηγμένες τεχνολογίες για τον μετασχηματισμό των ιστοσελίδων, ειδικά του ηλεκτρονικού εμπορίου, σε ένα εργαλείο για πολλαπλές συσκευές, αυξάνοντας την προσβασιμότητα και τη χρηστικότητα τους. Στη συνέχεια, πραγματοποιείται μια αξιολόγηση των PWA σχετικά με την προσβασιμότητα που προσφέρουν. Σε συνέχεια της έρευνας, εξετάζεται συστηματικά κάθε πτυχή της Βελτιστοποίησης Μηχανών Αναζήτησης (SEO), εξερευνώντας ευκαιρίες σε διάφορους τομείς κάτω από την ομπρέλα του ηλεκτρονικού εμπορίου, συμπεριλαμβανομένων του λιανικού ηλεκτρονικού εμπορίου και των τομέων που παρέχουν υπηρεσίες, όπως οι πλατφόρμες κρατήσεων. Ξεκινώντας με την αναγνώριση των τεχνικών και τεχνολογιών SEO, εντοπίζονται εκείνες οι τεχνικές που είναι πιο αποτελεσματικές για την εξασφάλιση υψηλότερης κατάταξης στα αποτελέσματα αναζήτησης και αυξημένης οργανικής επισκεψιμότητας. Τέλος, εξετάζεται η τεχνολογία Επιταχυνόμενες Σελίδες για Κινητά (AMP) και πώς συμβάλλει στην αύξηση της επισκεψιμότητας σε πλατφόρμες ηλεκτρονικού εμπορίου.

Κατά την έρευνα των Μεγάλων Γλωσσικών Μοντέλων (LLMs), ερευνήθηκε αρχικά η Επεξεργασία Φυσικής Γλώσσας (NLP) και η ικανότητά της να κατανοεί την ανθρώπινη γλώσσα, την αρχιτεκτονική του Generative Pre-trained Transformer (GPT) και την καινοτόμο χρήση των transformers και των μηχανισμών αυτο-προσοχής για την επεξεργασία ακολουθιών εισόδου. Ερευνήθηκε παράλληλα η φάση της προ-εκπαίδευσης που περιλαμβάνει ένα δισεκατομμύριο παραμέτρους που χρησιμοποιούνται κατά την προ-εκπαίδευση των μοντέλων GPT, καθώς και τη δυνατότητά τους να υποστούν λεπτομερή ρύθμιση (FT) για εργασίες που αφορούν συγκεκριμένους τομείς. Στη συνέχεια, η έρευνα ειδικεύτηκε στα μοντέλα GPT-3.5, GPT-4 και LLaMA-2, εξετάζοντας την ενσωμάτωσή τους στο ηλεκτρονικό εμπόριο και τις πιθανές βελτιώσεις, συμπεριλαμβανομένης της αυτοματοποίησης, που θα μπορούσαν να προσφέρουν για τη βελτίωση των λειτουργιών και των πελατοκεντρικών πτυχών του ηλεκτρονικού εμπορίου. Αξίζει να σημειωθεί ότι η παρούσα έρευνα υπερβαίνει τα LLMs, καθώς εξετάστηκαν λεπτομερώς διάσημα μοντέλα NLP, όπως το BERT και το RoBERTa. Επιπλέον, εξετάστηκαν αλγόριθμοι μη εποπτευόμενης και εποπτευόμενης μάθησης, όπως οι ομαδοποιήσεις k-means, το φίλτράρισμα βασισμένο στο περιεχόμενο (CBF), οι ιεραρχικές ομαδοποιήσεις, καθώς και οι αλγόριθμοι λογιστικής παλινδρόμησης και νευρωνικών δικτύων. Για τους σκοπούς της έρευνας, δημιουργήθηκαν εφαρμογές Chrome και APIs βασισμένα σε Flask χρησιμοποιώντας Python και JavaScript. Τα προαναφερθέντα μοντέλα υπέστησαν λεπτομερή ρύθμιση (FT) για συγκεκριμένους τομείς, παρέχοντας αξιολογικά αποτελέσματα για την ενσωμάτωση των LLMs και NLP στον τομέα

του ηλεκτρονικού εμπορίου. Η έρευνα εστίασε σε κορυφαία θέματα, συμπεριλαμβανομένης της ανάλυσης συναισθημάτων, συστημάτων σύστασης, συστημάτων απόφασης βιώσιμων αγορών και του μοντελοποιητικού χειρισμού αποχώρησης. Αυτή η σφαιρική εξέταση είχε ως στόχο την αποκάλυψη όλων εκείνων πρακτικών εφαρμογών και του οφέλους από τη χρήση προηγμένων μοντέλων γλώσσας για τη βελτιστοποίηση του ηλεκτρονικού εμπορίου.

Κατά την εξέταση της προσβασιμότητας στον Ιστό, μέσω μιας κριτικής ανασκόπησης των Οδηγιών Προσβασιμότητας Περιεχομένου του Ιστού (WCAG), αναγνωρίστηκαν και παρουσιάστηκαν λεπτομερώς κάθε αρχή σχεδίασης, οδηγία, σημείο ελέγχου και κριτήριο επιτυχίας. Ως αποτέλεσμα της έρευνας, εντοπίστηκαν περιορισμοί στα WCAG τόσο όσον αφορά τις οδηγίες προσβασιμότητας, όσο και στις προσπάθειες των προγραμματιστών ιστού και τα εργαλεία αυτοματοποιημένου ελέγχου. Καθώς η έρευνα προχωρούσε, μια αναδυόμενη τεχνολογία εφαρμογής ιστού που είναι γνωστή ως Προοδευτικές Εφαρμογές Ιστού (PWAs) ερευνήθηκε σε βάθος. Οι PWAs βασίζονται σε έναν συνδυασμό προηγμένων τεχνολογιών, όπως οι υπηρεσίες service workers, η αρχιτεκτονική του κελύφους εφαρμογής, το μανιφέστο της εφαρμογής ιστού και η αποθήκευση προσωρινής μνήμης για λειτουργία εκτός σύνδεσης. Κατά τη διάρκεια μιας παράλληλης κριτικής αξιολόγησης, εξετάστηκαν τα πλεονεκτήματα και οι περιορισμοί της υιοθέτησης αυτής της καινοτόμου τεχνολογίας. Συνδυάζοντας τις γνώσεις που αποκομίστηκαν σχετικά με την Προσβασιμότητα και τις PWAs, επιλέχθηκε ένα δείγμα ιστοτόπων και πλατφορμών ηλεκτρονικού εμπορίου που χρησιμοποιούν την τεχνολογία αυτή. Χρησιμοποιώντας δημοφιλή εργαλεία αξιολόγησης προσβασιμότητας και εργαλεία ελέγχου, διενεργήθηκε χειροκίνητη συγκριτική ανάλυση έναντι των αντίστοιχων μη-PWA. Μετά τη συλλογή δειγμάτων και την περιγραφική ανάλυση με τη χρήση Jupyter και Python, η έρευνα κατέληξε σε πολύτιμα αποτελέσματα σχετικά με την προσβασιμότητα των PWAs σε σύγκριση με τους ιστότοπους που δεν χρησιμοποιούν PWA. Αυτή η μελέτη υπογραμμίζει τον απαραίτητο ρόλο της προσβασιμότητας στον ιστό.

Κατά τη διάρκεια της συστηματικής ανασκόπησης της βιβλιογραφίας (SLR) στη Βελτιστοποίηση Μηχανών Αναζήτησης (SEO), τονίστηκε η αναγκαιότητα των τεχνικών και τεχνολογιών SEO τόσο για τη σωστή παρουσία των ηλεκτρονικών καταστημάτων στον Παγκόσμιο Ιστό όσο και για υψηλότερες κατατάξεις στα αποτελέσματα αναζήτησης οργανικού τύπου (SERPs). Η παρούσα έρευνα ωστόσο δεν περιορίζεται μόνο στις υπάρχουσες Τεχνικές SEO. Αντίθετα, προτείνεται η χρήση υπαρχουσών τεχνολογιών ιστού για την ενίσχυση του SEO. Για τους σκοπούς της έρευνας, δημιουργήθηκαν πρωτότυπα εργαλεία σε PHP και Python με χαρακτηριστικά παρόμοια με αυτά ενός crawler. Αυτά τα εργαλεία αναλύουν στοχευμένα ιστοσελίδες, εξάγοντας πολύτιμα στοιχεία σχετικά με τις τεχνικές SEO που χρησιμοποιούνται σε αυτές. Για να καλυφθούν οι τρέχουσες ανάγκες στο SEO, τα συγκεκριμένα εργαλεία συνδυάστηκαν με APIs, επιτρέποντάς τους να ανακτούν επιπλέον δεδομένα, όπως ο αριθμός των backlinks, η εξουσία του τομέα (DA), λέξεις-κλειδιά, θέσεις στις αναζητήσεις, σχεδιασμό για πολλαπλές συσκευές και ταχύτητα της ιστοσελίδας. Επιπλέον, μέσω της ενσωμάτωσης της μηχανικής μάθησης και αλγορίθμων πρόβλεψης στα λογισμικά αυτά, δημιουργήθηκε ένα προ-εκπαιδευμένο μοντέλο. Αυτό το μοντέλο διευκόλυνε την ακριβή πρόβλεψη του απαραίτητου αριθμού backlinks και της εξουσίας του τομέα (DA) που πρέπει να έχει μια ιστοσελίδα για να κερδίσει την κατάταξη στην πρώτη σελίδα των αποτελεσμάτων αναζήτησης, λαμβάνοντας υπόψη το ανταγωνιστικό της περιβάλλον. Τα εργαλεία που δημιουργήθηκαν για τις ανάγκες της έρευνας έχουν δημοσιευτεί ως ανοικτού κώδικα στο GitHub, παρέχοντας σε μη-ειδικευμένες ομάδες SEO και ιδιοκτήτες επιχειρήσεων τη δυνατότητα να υλοποιήσουν στρατηγικές SEO και να παρατηρήσουν τα αποτελέσματα μέσω των συστάσεων του λογισμικού. Για να επιβεβαιωθεί η αποτελεσματικότητα των εργαλείων, καθώς και για να βρεθούν ποιες από τις τεχνικές SEO είναι πιο αποτελεσματικές, διεξήχθησαν μελέτες περιπτώσεων σε συγκεκριμένους τομείς ηλεκτρονικού εμπορίου.

Στο τελευταίο στάδιο της έρευνας, παρατηρήθηκε η άνοδος της βιομηχανίας κινητών συσκευών, που ώθησε τους διαχειριστές ιστού να εξερευνήσουν εναλλακτικές μεθόδους για την

ανάπτυξη ιστοσελίδων που είναι περισσότερο φιλικές προς τις κινητές συσκευές. Μια τεχνολογία που κέρδισε το ενδιαφέρον των διαχειριστών ιστού ήταν οι Επιταχυνόμενες Σελίδες Κινητών της Google (AMP). Χρησιμοποιώντας υπάρχουσες τεχνολογίες και επιβάλλοντας κάποιους περιορισμούς, οι AMP είχαν ως στόχο τη δημιουργία προ-φορτωμένων ιστοσελίδων με πλεονεκτήματα όσον αφορά τον χρόνο φόρτωσης. Αυτές οι σελίδες προσφέρουν όχι μόνο μείωση του μεγέθους της σελίδας, αλλά και βελτίωση των κατατάξεων στα αποτελέσματα αναζήτησης, με αποτέλεσμα την αυξημένη οργανική επισκεψιμότητα λόγω της ταχύτητάς τους και της φιλικότητάς τους προς τον χρήστη. Σε αυτήν τη μελέτη, για να εξακριβωθεί εάν οι AMP πράγματι παρέχουν τα υποσχόμενα οφέλη σε ταχύτητα και κατάταξη, πραγματοποιήθηκε συγκριτική ανάλυση στις σελίδες ενός ιστοτόπου ηλεκτρονικού εμπορίου σε σχέση με τις αντίστοιχες σελίδες του ίδιου ιστοτόπου που χρησιμοποιούσαν την τεχνολογία AMP.

*Λέξεις Κλειδιά:* Βελτιστοποίηση Μηχανών Αναζήτησης, Επεξεργασία Φυσικής Γλώσσας, Επιταχυνόμενες Σελίδες για Κινητές Συσκευές, Ηλεκτρονικό Εμπόριο, Ικανοποίηση Πελατών, Λεπτομερής Ρύθμιση Μοντέλων, Μάθηση Μοντέλων με Ελάχιστα Δεδομένα, Μεγάλα Γλωσσικά Μοντέλα, Μη Εποπτευόμενη Μάθηση, Μηχανική Μάθηση, Μοντελοποιητικός Χειρισμός Αποχώρησης, Οργανική Κίνηση, Προ-εκπαιδευμένα Μοντέλα, Προοδευτικές Εφαρμογές Ιστού, Προσβασιμότητα Ιστού, Συστήματα Σύστασης

*Dedicated to my wife and our little John for their unconditional love, support,  
and encouragement.*



## Preface

This thesis is submitted in fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Peloponnese. The research presented here was conducted at the University of Peloponnese under the supervision of Professor Nikolaos Tselikas.

The main subject of this thesis is to maximize web usability, user experience, accessibility, and web traffic in E-commerce by integrating software, Machine Learning (ML), Large Language Models (LLMs), Natural Language Processing (NLP), and Search Engine Optimization (SEO) techniques. This thesis is based on a collection of twelve papers published in international journals and conferences, which have been adapted to fit into four main chapters.

The main chapters are preceded by an introductory chapter that establishes their interrelation, provides background information on the study's significance, and includes a summary of each. Finally, the concluding chapter presents the main points and prospects of this work, bringing the thesis to a close.

## Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Nikolaos Tselikas, whose contribution has been invaluable. Thank you for your guidance, **patience**, careful supervision, and encouragement throughout the years.

I am also deeply grateful to Professor Costas Vassilakis and Professor Christos Tryfonopoulos for their support and invaluable feedback at every stage of my research journey.

I would also like to express my heartfelt gratitude to my family for their understanding, encouragement, and unwavering support. They have consistently been my greatest pillars throughout this entire process, urging me to persist when frustration and other challenges threatened my motivation. I cannot thank them enough for their love, assistance, and encouragement.

*Konstantinos I. Roumeliotis*



## List of Papers

### Journals

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Precision-Driven Product Recommendations Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems. Software 2024. <https://doi.org/10.3390/software3010004>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. Natural Language Processing Journal 2024, 6, 100056. <https://doi.org/10.1016/j.nlp.2024.100056>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Unveiling Sustainability in Ecommerce: GPT-Powered Software for Identifying Sustainable Product Features. Sustainability 2023, 15, 12015. <https://doi.org/10.3390/su151512015>

Roumeliotis, K.I.; Tselikas, N.D. A Machine Learning Python-Based Search Engine Optimization Audit Software. Informatics 2023, 10, 68. <https://doi.org/10.3390/informatics10030068>

Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet 2023, 15, 192. <https://doi.org/10.3390/fi15060192>

Roumeliotis, K.I.; Tselikas, N.D. Evaluating Progressive Web App Accessibility for People with Disabilities. Network 2022, 2, 350-369. <https://doi.org/10.3390/network2020022>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Airlines' Sustainability Study Based on Search Engine Optimization Techniques and Technologies. Sustainability 2022, 14, 11225. <https://doi.org/10.3390/su141811225>

Roumeliotis, K.I.; Tselikas, N.D.; Tryfonopoulos, C. Greek Hotels' Web Traffic: A Comparative Study Based on Search Engine Optimization Techniques and Technologies. Digital 2022, 2, 379-400. <https://doi.org/10.3390/digital2030021>

Roumeliotis, K.I.; Tselikas, N.D. An Effective SEO Techniques and Technologies Guide-map. Journal of Web Engineering 2022, 21, 5, 1603–1650. <https://doi.org/10.13052/jwe1540-9589.21510>

### Conferences

Roumeliotis, K.I.; Tselikas, N.D. Churn Modeling in E-commerce: Leveraging the Power of Logistic Regression and Neural Networks. Conference on Business Intelligence and Modelling. Springer Proceedings in Business and Economics. (Acceptance Letter. Pending DOI)

Roumeliotis, K.I.; Tselikas, N.D. Search Engine Optimization Techniques: The Story of an Old-Fashioned Website. International Conference on Business Intelligence and Modelling. Springer Proceedings in Business and Economics. Springer, Cham 2021, 47–55. [https://doi.org/10.1007/978-3-030-57065-1\\_4](https://doi.org/10.1007/978-3-030-57065-1_4)

Roumeliotis, K.I.; Tselikas, N.D. Accelerated Mobile Pages: A Comparative Study. International Conference on Business Intelligence and Modelling. Springer Proceedings in Business and Economics. Springer, Cham 2021, 57–63. [https://doi.org/10.1007/978-3-030-57065-1\\_5](https://doi.org/10.1007/978-3-030-57065-1_5)

# Table of Contents

ABSTRACT .....	4
ΠΕΡΙΛΗΨΗ .....	6
PREFACE.....	11
ACKNOWLEDGEMENTS .....	11
LIST OF PAPERS .....	13
TABLE OF CONTENTS .....	14
TABLE OF FIGURES .....	18
TABLE OF TABLES .....	21
<b>1. INTRODUCTION .....</b>	<b>23</b>
1.1. THE ROLE AND NECESSITY OF AI IN E-COMMERCE .....	23
1.1.1. <i>Enhancing the E-Commerce Experience through AI</i> .....	23
1.1.2. <i>Key Competitive Advantages Using AI</i> .....	24
1.2. THE IMPERATIVE OF WEB ACCESSIBILITY IN E-COMMERCE .....	24
1.2.1. <i>Beyond Moral and Legal Obligations</i> .....	25
1.3. PROGRESSIVE WEB APPS: BRIDGING ACCESSIBILITY AND ENGAGEMENT IN E-COMMERCE .....	25
1.3.1. <i>Defining Progressive Web Apps</i> .....	25
1.3.2. <i>Core Features of PWAs</i> .....	25
1.3.3. <i>Enhancing Engagement and Conversion Rates</i> .....	26
1.4. SEARCH ENGINE OPTIMIZATION (SEO) FOR ELEVATED E-COMMERCE VISIBILITY AND GROWTH.....	26
1.4.1. <i>The Relationship Between Rankings, Traffic, and Business Growth</i> .....	26
1.4.2. <i>Using SEO for Shaping E-Commerce Success</i> .....	26
1.5. REVOLUTIONIZING MOBILE E-COMMERCE WITH ACCELERATED MOBILE PAGES (AMP) .....	27
1.5.1. <i>AMP and Its Relevance to Mobile E-Commerce</i> .....	27
1.5.2. <i>Benefits of AMP for E-Commerce Websites</i> .....	27
1.5.3. <i>Potential of AMP to Increase Visibility and Engagement in E-Commerce Sales Funnel</i> .....	28
1.6. COLLECTIVE CONTRIBUTION: NECESSITY FOR E-COMMERCE GROWTH AND SUSTAINABILITY .....	28
<b>2. LARGE LANGUAGE MODELS (LLMS) AND NATURAL LANGUAGE PROCESSING (NLP) IN E-COMMERCE ...32</b>	
2.1. INTRODUCTION.....	32
2.2. NATURAL LANGUAGE PROCESSING (NLP).....	33
2.3. GENERATIVE PRE-TRAINED TRANSFORMER (GPT) ARCHITECTURE .....	33
2.4. PRE-TRAINING PHASE AND SUPERVISED FINE-TUNING .....	33
2.5. PROMPTS IN GPT MODELS.....	34
2.6. LLAMA-2 MODEL.....	35
2.6.1. <i>LLaMA-2 Models and Versions</i> .....	35
2.6.2. <i>Pretraining Phase</i> .....	35
2.6.3. <i>LLaMA-2 Fine-tuning</i> .....	36
2.7. LLMS IN E-COMMERCE: A COMPARATIVE ANALYSIS OF GPT AND LLAMA MODELS IN PRODUCT REVIEW EVALUATION ..	36
2.7.1. <i>Section Summary and Significance of the Study</i> .....	36
2.7.2. <i>Literature review</i> .....	38
2.7.3. <i>Research methodology</i> .....	40
2.7.4. <i>Comparative analysis and models' evaluation</i> .....	46
2.7.5. <i>Research findings</i> .....	50
2.7.6. <i>Implications and directions for future research</i> .....	55
2.7.7. <i>Section Conclusion</i> .....	56
2.8. PRECISION-DRIVEN PRODUCT RECOMMENDATIONS SOFTWARE: UNSUPERVISED MODELS, EVALUATED BY GPT-4 LLM FOR ENHANCED RECOMMENDER SYSTEMS .....	56
2.8.1. <i>Section Summary and Significance of the Study</i> .....	56
2.8.2. <i>Literature Review</i> .....	57

2.8.3.	<i>Materials and Methods</i> .....	60
2.8.4.	<i>Research Results</i> .....	64
2.8.5.	<i>Recommender Systems Evaluation and Discussion</i> .....	68
2.8.6.	<i>Section Conclusion</i> .....	70
2.9.	GPT-POWERED SOFTWARE FOR IDENTIFYING SUSTAINABLE PRODUCT FEATURES .....	71
2.9.1.	<i>Section Summary and Significance of the Study</i> .....	71
2.9.2.	<i>Sustainability Literature Review</i> .....	73
2.9.3.	<i>Materials and Methods</i> .....	77
2.9.4.	<i>Experimental Setup and Results</i> .....	81
2.9.5.	<i>Discussion</i> .....	86
2.9.6.	<i>Section Conclusion</i> .....	88
2.10.	CHURN MODELING IN E-COMMERCE: LEVERAGING THE POWER OF LOGISTIC REGRESSION AND NEURAL NETWORKS .....	88
2.10.1.	<i>Section Summary and Significance of the Study</i> .....	88
2.10.2.	<i>Literature Review</i> .....	88
2.10.3.	<i>Methodology</i> .....	90
2.10.4.	<i>Results and Discussion</i> .....	91
2.10.5.	<i>Performance Comparison of Logistic Regression and Neural Networks</i> .....	92
2.10.6.	<i>Section Conclusion</i> .....	94
<b>3.</b>	<b>WEB ACCESSIBILITY AND PROGRESSIVE WEB APPS (PWA) .....</b>	<b>95</b>
3.1.	INTRODUCTION.....	95
3.2.	MATERIALS AND METHODS.....	95
3.2.1.	<i>Research Limitations</i> .....	96
3.3.	PROGRESSIVE WEB APPS (PWA).....	96
3.3.1.	<i>Service Workers</i> .....	97
3.3.2.	<i>Caching Storage Application Programming Interface (API) and Offline Functionality</i> .....	98
3.3.3.	<i>Web App Manifest</i> .....	99
3.3.4.	<i>Push API and Notifications</i> .....	100
3.3.5.	<i>Application Shell (App Shell) Architecture</i> .....	100
3.3.6.	<i>PWA Critical View</i> .....	100
3.4.	WEB ACCESSIBILITY .....	102
3.4.1.	<i>Guidelines and Regulations</i> .....	102
3.4.2.	<i>Web Accessibility Critical View</i> .....	105
3.5.	RESULTS.....	106
3.6.	SECTION CONCLUSION .....	112
<b>4.</b>	<b>SEARCH ENGINE OPTIMIZATION (SEO) IN E-COMMERCE .....</b>	<b>113</b>
4.1.	INTRODUCTION.....	113
4.2.	ON-PAGE SEO TECHNIQUES .....	115
4.2.1.	<i>Copywriting and Keyword Optimization</i> .....	115
4.2.2.	<i>Crawlable URLs – SEO Friendly URLs</i> .....	118
4.2.3.	<i>Descriptive Meta Tags</i> .....	119
4.2.4.	<i>Heading Tags</i> .....	119
4.2.5.	<i>Optimizing Images for Search Engines</i> .....	119
4.2.6.	<i>Tags on Images and URLs</i> .....	120
4.2.7.	<i>Rating and Review</i> .....	120
4.2.8.	<i>Sitemaps and RSS Feed</i> .....	122
4.2.9.	<i>Robots.txt</i> .....	122
4.2.10.	<i>Breadcrumbs</i> .....	123
4.3.	OFF-PAGE SEO TECHNIQUES .....	125
4.4.	TECHNOLOGIES PROMOTING SEO.....	126
4.4.1.	<i>Responsive Design - Mobile Friendliness</i> .....	126
4.4.2.	<i>Website Speed</i> .....	127
4.4.3.	<i>Serving Data over HTTPS</i> .....	127
4.4.4.	<i>Structured Data and Rich Snippets</i> .....	128
4.4.5.	<i>Accelerated Mobile Pages Project (AMP)</i> .....	130
4.4.6.	<i>Apache Deflate and Gzip</i> .....	131

4.4.7.	<i>Page Caching</i> .....	131
4.4.8.	<i>Minify Resources (HTML, CSS, and JavaScript)</i> .....	131
4.4.9.	<i>Open Graph Protocol (OGP)</i> .....	132
4.4.10.	<i>Content Delivery Network (CDN) Links</i> .....	133
4.5.	BLACK HAT SEO TECHNIQUES.....	134
4.6.	PAID TRAFFIC.....	134
4.7.	A MACHINE LEARNING PYTHON-BASED SEARCH ENGINE OPTIMIZATION AUDIT SOFTWARE.....	135
4.7.1.	<i>Section Summary and Significance of the Study</i> .....	135
4.7.2.	<i>Materials and Methods</i> .....	136
4.7.3.	<i>Results</i> .....	143
4.7.4.	<i>Discussion</i> .....	147
4.7.5.	<i>Section Conclusion</i> .....	150
4.8.	SEO TECHNIQUES AND TECHNOLOGIES EVALUATION.....	151
4.8.1.	<i>Section Summary and Significance of the Study</i> .....	151
4.8.2.	<i>Materials and Methods</i> .....	151
4.8.3.	<i>Research Findings</i> .....	152
4.8.4.	<i>Discussion and Future Work</i> .....	154
4.8.5.	<i>Section Conclusion</i> .....	155
4.9.	AIRLINES' SUSTAINABILITY STUDY BASED ON SEARCH ENGINE OPTIMIZATION TECHNIQUES AND TECHNOLOGIES.....	155
4.9.1.	<i>Section Summary and Significance of the Study</i> .....	155
4.9.2.	<i>Materials and Methods</i> .....	155
4.9.3.	<i>Results</i> .....	160
4.9.4.	<i>Discussion</i> .....	169
4.9.5.	<i>Section Conclusion</i> .....	170
4.10.	GREEK HOTELS' WEB TRAFFIC: A COMPARATIVE STUDY BASED ON SEARCH ENGINE OPTIMIZATION TECHNIQUES AND TECHNOLOGIES.....	171
4.10.1.	<i>Section Summary and Significance of the Study</i> .....	171
4.10.2.	<i>Materials and Methods</i> .....	171
4.10.3.	<i>Results</i> .....	175
4.10.4.	<i>Discussion</i> .....	182
4.10.5.	<i>Section Conclusion</i> .....	183
4.11.	THE SEARCH ENGINE OPTIMIZATION STORY OF AN OLD-FASHIONED E-COMMERCE.....	184
4.11.1.	<i>Section Summary and Significance of the Study</i> .....	184
4.11.2.	<i>Materials and Methods</i> .....	184
4.11.3.	<i>Search Engine Optimization Results</i> .....	184
4.11.4.	<i>Section Research Findings and Discussion</i> .....	186
4.11.5.	<i>Section Conclusion</i> .....	187
<b>5.</b>	<b>THE ACCELERATED MOBILE PAGES (AMP) AND ITS CONTRIBUTION TO WEB TRAFFIC IN THE REALM OF E-COMMERCE</b> .....	<b>188</b>
5.1.	SECTION SUMMARY AND SIGNIFICANCE OF THE STUDY.....	188
5.2.	AMP USAGE AND RESTRICTIONS.....	188
5.2.1.	<i>How AMP works</i> .....	188
5.2.2.	<i>AMP Restrictions and Components</i> .....	189
5.3.	AMP IMPACT TO A REAL-LIFE E-COMMERCE.....	190
5.3.1.	<i>AMP Speed Performance</i> .....	190
5.3.2.	<i>Comparing Organic Traffic with AMP Traffic</i> .....	190
5.4.	SECTION CONCLUSION.....	191
	<b>CONCLUSION</b> .....	<b>192</b>
	<b>BIBLIOGRAPHY</b> .....	<b>194</b>



# Table of Figures

<b>FIGURE 1.</b> MODEL-AGNOSTIC PROMPT .....	42
<b>FIGURE 2.</b> FINE-TUNING JSONL SAMPLES .....	44
<b>FIGURE 3.</b> MODEL-AGNOSTIC PROMPT FOR IDENTIFYING LEXICAL ELEMENTS IN PRODUCT REVIEWS.....	46
<b>FIGURE 4.</b> MAE FOR BASE MODELS.....	48
<b>FIGURE 5.</b> MAE FOR FINE-TUNED MODELS.....	49
<b>FIGURE 6.</b> MODEL-WISE VARIABILITY IN ACCURACY PERCENTAGE OF MODEL PREDICTIONS .....	54
<b>FIGURE 7.</b> GPT PROMPT FOR EVALUATING PRODUCT RECOMMENDATIONS.....	63
<b>FIGURE 8.</b> TRAINING AND PRODUCT RECOMMENDATIONS USING THE FLASK API .....	64
<b>FIGURE 9.</b> GPT-4 EVALUATIONS OF UNSUPERVISED LEARNING TRAINED MODELS.....	65
<b>FIGURE 10.</b> PERCENTAGE OF USE OF SUSTAINABLE FEATURES ACROSS PRODUCTS.....	80
<b>FIGURE 11.</b> NUMBER OF OCCURRENCES OF SUSTAINABLE FEATURES BY MARKETPLACE .....	80
<b>FIGURE 12.</b> PROMPT TO OBTAIN THE APPROPRIATE RESPONSE FROM THE GPT MODEL .....	82
<b>FIGURE 13.</b> SCREENSHOT OF A PRODUCT PAGE BEFORE RECEIVING THE RESPONSE FROM THE CHROME APP AND GPT MODEL .....	83
<b>FIGURE 14.</b> THE SAME PRODUCT PAGE DISPLAYING MODIFICATIONS AFTER THE RESPONSE GENERATED BY THE CHROME APP AND GPT MODEL.....	83
<b>FIGURE 15.</b> THE NON-SUSTAINABLE PRODUCT PAGE DISPLAYING MODIFICATIONS AFTER THE RESPONSE GENERATED BY THE CHROME APP AND GPT MODEL .....	84
<b>FIGURE 16.</b> CHURN DISTRIBUTION BY PREFERREDLOGINDEVICE AND SATISFACTIONSORE .....	92
<b>FIGURE 17.</b> CONFUSION MATRICES FOR PREDICTIONS MADE BY THE TRAINED MODELS .....	93
<b>FIGURE 18.</b> OVERALL STEP-BY-STEP REPRESENTATION OF THE PROPOSED METHODOLOGY, REFLECTING THE STAGES FOR EVALUATING PWAS’ ACCESSIBILITY.....	96
<b>FIGURE 19.</b> SERVICE WORKER’S LIFECYCLE .....	98
<b>FIGURE 20.</b> CACHING MECHANISM.....	99
<b>FIGURE 21.</b> IMPORT REQUIRED IN PYTHON LIBRARIES .....	108
<b>FIGURE 22.</b> READ DATA INTO PANDAS DATA FRAME .....	108
<b>FIGURE 23.</b> PRESENTING OUTCOMES INTO PLOTS—EQUAL ACCESS BY IBM (T9).....	110
<b>FIGURE 24.</b> EQUAL ACCESS BY IBM (T9) ERRORS FOR NEW TECHNOLOGY (PWA) WEBSITES.....	110
<b>FIGURE 25.</b> EQUAL ACCESS BY IBM (T9) ERRORS FOR OLDER TECHNOLOGY (NON-PWA) WEBSITES.....	110
<b>FIGURE 26.</b> PWA VS. NON-PWA INDUSTRY PLOTS—AXE DEVTOOLS (T7).....	111
<b>FIGURE 27.</b> PRESENTING OUTCOMES PER INDUSTRY INTO PLOTS (RED: OLDER TECHNOLOGY, BLUE: NEW TECHNOLOGY)—AXE DEVTOOLS (T7) .....	111
<b>FIGURE 28.</b> DIFFERENCES BETWEEN SEO FRIENDLY AND NOT SEO FRIENDLY URL .....	118
<b>FIGURE 29.</b> SOURCE CODE FROM AN .HTACCESS FILE.....	119
<b>FIGURE 30.</b> APPROPRIATE <META> DESCRIPTION AND <META> KEYWORDS .....	119
<b>FIGURE 31.</b> IMAGE SEO FILENAME & ALT TAG .....	120
<b>FIGURE 32.</b> LINK SEO & TITLE TAG.....	120
<b>FIGURE 33.</b> STRUCTURED DATA RATINGS AND REVIEWS – A) MICRODATA AND B) JSON-LD .....	121
<b>FIGURE 34.</b> RATINGS, REVIEWS AND FEEDBACK ON SEARCH RESULTS .....	122
<b>FIGURE 35.</b> SITEMAP EXAMPLE .....	122
<b>FIGURE 36.</b> RSS ARTICLE EXAMPLE.....	122
<b>FIGURE 37.</b> ROBOTS.TXT AND ROBOT META .....	123
<b>FIGURE 38.</b> BREADCRUMBS.....	123
<b>FIGURE 39.</b> USING BOOTSTRAP TO VISUAL DESIGN BREADCRUMBS.....	123
<b>FIGURE 40.</b> STRUCTURED DATA – A) MICRODATA, B) RDFa AND C) JSON-LD TO IMPROVE BREADCRUMBS.....	124
<b>FIGURE 41.</b> BREADCRUMB SHOWS WEBSITE HIERARCHY AND CATEGORIZATION .....	124
<b>FIGURE 42.</b> DA100 AND DA40 CLOTHING WEBSITES POINT BACKLINK TO RELATED CLOTHING WEBSITE AND NOT RELATED PERFUME WEBSITE .....	125
<b>FIGURE 43.</b> GOOGLE ANALYTICS - BOUNCE RATE - SESSION DURATION.....	126
<b>FIGURE 44.</b> RESPONSIVE DESIGN FOR MULTIPLE DEVICES – BOOTSTRAP .....	127
<b>FIGURE 45.</b> STRUCTURED DATA ON PRODUCT A) MICRODATA B) RDFa C) JSON-LD.....	129
<b>FIGURE 46.</b> SEARCH RESULTS WITHOUT AND WITH AMP.....	130
<b>FIGURE 47.</b> CANONICAL AND AMP HTML LINK ON HEADER.....	131
<b>FIGURE 48.</b> FILESIZE UNCOMPRESSED & COMPRESSED JQuery .....	132
<b>FIGURE 49.</b> THE OPENGRAPH PROTOCOL EXAMPLE.....	133

<b>FIGURE 50.</b> FACEBOOK PAGE POST USING OG, EARN VISITOR TO WEBSITE.....	133
<b>FIGURE 51.</b> PARTIAL ADOPTION OF CDN (BOOTSTRAP, JQUERY AND FONT AWESOME LIBRARIES).....	134
<b>FIGURE 52.</b> TRAIN MODEL PHASE: DA TO BACKLINKS AND LINKING DOMAIN.....	142
<b>FIGURE 53.</b> SCATTERPLOTS FOR LINKINGDOMAINS AND BACKLINKS.....	142
<b>FIGURE 54.</b> MAKE PREDICTIONS PHASE: DA TO BACKLINKS AND LINKING DOMAIN.....	143
<b>FIGURE 55.</b> GOOGLE ANALYTICS. TRAFFIC TREND FROM 26 FEBRUARY 2023 TO 20 MAY 2023.....	147
<b>FIGURE 56.</b> SEO TECHNIQUES USAGE SCORE AMONG TOP-LISTED WEBSITES (%). .....	152
<b>FIGURE 57.</b> OVERALL STEP-BY-STEP REPRESENTATION OF THE PROPOSED METHODOLOGY, REFLECTING THE STAGES FOR COMPARING SEO TECHNIQUES AND SEO METRICS TO WEB TRAFFIC. ....	156
<b>FIGURE 58.</b> PHP-BASED SEO SOFTWARE.....	158
<b>FIGURE 59.</b> TESTING RESEARCH MODEL.....	161
<b>FIGURE 60.</b> SCATTERPLOT OF DA BY WEB TRAFFIC.....	167
<b>FIGURE 61.</b> SCATTERPLOT OF KEYWORDS BY WEB TRAFFIC.....	167
<b>FIGURE 62.</b> SCATTERPLOT OF WEBSITE SPEED BY WEB TRAFFIC.....	168
<b>FIGURE 63.</b> SCATTERPLOT OF BACKLINKS BY WEB TRAFFIC.....	168
<b>FIGURE 64.</b> USING FUZZY COGNITIVE MAP TO LOCATE THE CORRELATIONS BETWEEN SEO TECHNIQUES/METRICS AND WEB TRAFFIC. .....	169
<b>FIGURE 65.</b> SEO TECHNIQUES ADOPTION PERCENTAGE.....	170
<b>FIGURE 66.</b> PHP-BASED SEO SOFTWARE.....	173
<b>FIGURE 67.</b> SCATTERPLOT OF EACH METRIC BY WEB TRAFFIC.....	181
<b>FIGURE 68.</b> USING FUZZY COGNITIVE MAP TO LOCATE THE CORRELATIONS BETWEEN SEO TECHNIQUES/METRICS AND WEB TRAFFIC. .....	182
<b>FIGURE 69.</b> SEO TECHNIQUES ADOPTION PERCENTAGE.....	183
<b>FIGURE 70.</b> ORGANIC TRAFFIC PER DAY BEFORE SEO IMPLEMENTATION - STATISTICS EXPORTED FROM GOOGLE ANALYTICS .....	184
<b>FIGURE 71.</b> STRUCTURED DATA RESULTS.....	186
<b>FIGURE 72.</b> INDEXING RESULTS BETWEEN META TAGS AND OG TAGS .....	186
<b>FIGURE 73.</b> ORGANIC TRAFFIC PER WEEK - STATISTICS EXPORTED FROM GOOGLE ANALYTICS .....	187
<b>FIGURE 74.</b> ORGANIC TRAFFIC PER DAY - STATISTICS EXPORTED FROM GOOGLE ANALYTICS .....	187
<b>FIGURE 75.</b> PROCEDURE OF CACHING.....	189
<b>FIGURE 76.</b> AMP IN-LINE CSS STYLING.....	189
<b>FIGURE 77.</b> AMP EXTERNAL JAVASCRIPT.....	189
<b>FIGURE 78.</b> AMP URL STRUCTURE.....	190
<b>FIGURE 79.</b> AMP HTML AND CANONICAL LINK TAG.....	190
<b>FIGURE 80.</b> AMP VS NON-AMP WEBPAGE SPEED.....	190
<b>FIGURE 81.</b> AMP VS ORGANIC TRAFFIC.....	191



# Table of Tables

<b>TABLE 1.</b> MODEL PERFORMANCE METRICS COMPARISON.....	46
<b>TABLE 2.</b> LEXICAL ELEMENTS DETECTED BY FINE-TUNED LLMs IN REVIEW BODIES.....	49
<b>TABLE 3.</b> EFFICIENCY OF UNSUPERVISED LEARNING-TRAINED MODELS IN RECOMMENDER SYSTEMS.....	65
<b>TABLE 4.</b> IN-CART PRODUCT RECOMMENDATIONS AND GPT MODEL EVALUATION .....	66
<b>TABLE 5.</b> DESCRIPTIVE STATISTICS FOR THE SIMILARITY SCORES IN EACH RECOMMENDATION SYSTEM .....	69
<b>TABLE 6.</b> ACCURACY OF THE PREDICTIONS MADE BY THE LOGISTIC REGRESSION-TRAINED MODEL.....	92
<b>TABLE 7.</b> ACCURACY OF THE PREDICTIONS MADE BY THE NEURAL NETWORK-TRAINED MODEL.....	93
<b>TABLE 8.</b> WEB MANIFEST PROPERTIES.....	99
<b>TABLE 9.</b> ADVANTAGES AND DISADVANTAGES OF PWAs.....	100
<b>TABLE 10.</b> WCAG DESIGN PRINCIPLES, GUIDELINES, AND SUCCESS CRITERIA.....	104
<b>TABLE 11.</b> ACCESSIBILITY COMPLIANCE ERRORS FOR PWA WEBSITES .....	107
<b>TABLE 12.</b> ACCESSIBILITY COMPLIANCE ERRORS FOR NON-PWA WEBSITES.....	107
<b>TABLE 13.</b> DESCRIPTIVE ANALYSIS FOR NEW TECHNOLOGY (PWA)—NEW_TECH.DESCRIBE().....	109
<b>TABLE 14.</b> DESCRIPTIVE ANALYSIS FOR OLDER TECHNOLOGY (NON-PWA)—OLD_TECH.DESCRIBE().....	109
<b>TABLE 15.</b> SEO TECHNIQUES AND TECHNOLOGIES SUMMARY TABLE.....	114
<b>TABLE 16.</b> KEYWORD LEVEL TYPES .....	116
<b>TABLE 17.</b> COPYWRITING STRUCTURE.....	117
<b>TABLE 18.</b> METHODS AND THEIR CORRESPONDING SEO TECHNIQUES.....	137
<b>TABLE 19.</b> SEO ANALYSIS FOR COMPETITORS’ AND USER’S WEBSITES (REPORT-SEO_COMPETITORS_TABLE.CSV).....	143
<b>TABLE 20.</b> SEO RECOMMENDATIONS BASED ON COMPETITION (REPORT-SUGGESTIONS_BASED_ON_COMPETITION.CSV).....	144
<b>TABLE 21.</b> SEO RECOMMENDATIONS BASED ON SEO TOOL (REPORT-SEO_TOOL_SUGGESTIONS.CSV).....	145
<b>TABLE 22.</b> AN ILLUSTRATION OF THE RESULTS IN THE SEOMIZED DATABASE .....	153
<b>TABLE 23.</b> WEBSITE ACHIEVEMENTS AGAINST 17 SEO TECHNIQUES .....	153
<b>TABLE 24.</b> DESCRIPTION OF THE EXAMINED WEB ANALYTICS METRICS AND SUGGESTED KPIs AND PERFORMANCE MEASUREMENTS.....	157
<b>TABLE 25.</b> DESCRIPTIVE STATISTICS ON SEO METRICS.....	161
<b>TABLE 26.</b> INDEPENDENT T-TEST ON THE META DESCRIPTION TECHNIQUE.....	162
<b>TABLE 27.</b> INDEPENDENT T-TEST ON THE IMAGE ALT TECHNIQUE.....	163
<b>TABLE 28.</b> T-TEST ON 15 SEO TECHNIQUES.....	163
<b>TABLE 29.</b> MODEL SUMMARY. STRENGTH OF THE RELATIONSHIP BETWEEN THE MODEL AND THE DEPENDENT VARIABLE.....	164
<b>TABLE 30.</b> ANOVA RESULTS. OVERALL SIGNIFICANCE OF THE MODEL.....	165
<b>TABLE 31.</b> COEFFICIENT RESULTS. HOW MUCH DEPENDENT VARIABLE IS EXPECTED TO INCREASE WHEN THAT INDEPENDENT VARIABLE INCREASES BY ONE.....	165
<b>TABLE 32.</b> DESCRIPTION OF THE EXAMINED WEB ANALYTICS METRICS AND SUGGESTED KPIs AND PERFORMANCE MEASUREMENTS.....	172
<b>TABLE 33.</b> DESCRIPTIVE STATISTICS ON SEO METRICS.....	176
<b>TABLE 34.</b> INDEPENDENT T-TEST ON THE META DESCRIPTION TECHNIQUE.....	177
<b>TABLE 35.</b> INDEPENDENT T-TEST ON THE ROBOTS.TXT TECHNIQUE.....	177
<b>TABLE 36.</b> T-TEST ON 15 SEO TECHNIQUES.....	178
<b>TABLE 37.</b> MODEL SUMMARY. STRENGTH OF THE RELATIONSHIP BETWEEN THE MODEL AND THE DEPENDENT VARIABLE.....	179
<b>TABLE 38.</b> ANOVA RESULTS. OVERALL SIGNIFICANCE OF THE MODEL.....	179
<b>TABLE 39.</b> COEFFICIENT RESULTS. RESULTS SHOW HOW MUCH DEPENDENT VARIABLE IS EXPECTED TO INCREASE WHEN THAT INDEPENDENT VARIABLE INCREASES BY ONE.....	180
<b>TABLE 40.</b> PROS AND CONS OF USING AMP PROJECT.....	191



# 1. Introduction

In today's digital landscape, e-commerce is experiencing rapid growth, driven by technological advancements and shifting consumer preferences. The traditional boundaries of physical commerce have disappeared, giving rise to an interconnected network of online marketplaces. Within this dynamic realm, businesses fiercely compete for attention, and consumers leverage technology to make purchases from their screens.

The competition within this marketplace is intense, demanding that businesses be agile, adaptable, and innovative to thrive. Unlike traditional commerce, e-commerce is a fluid environment where businesses navigate challenges through technological solutions and innovation.

Armed with digital knowledge, consumers now expect more than mere transactions; they seek engaging experiences where each click and interaction unfolds seamlessly. To secure not just a click, but a purchase and loyalty from customers, businesses must not only meet but surpass these expectations.

The key to success in this dynamic e-commerce market is technology and innovation. Businesses utilizing these tools to innovate and adapt stand a better chance of thriving amid competition and evolving consumer expectations.

This thesis delves into the intricate relationship between technology, innovation, and e-commerce. By scrutinizing and combining cutting-edge advancements, the aim is to uncover strategies that not only sustain present-day e-commerce but also propel it into the future—a future where e-commerce remains a dynamic stage for those who master technological innovation.

## *1.1. The Role and Necessity of AI in E-Commerce*

In the dynamic realm of e-commerce, the rise of Artificial Intelligence (AI) emerges as a transformative influence, redefining the dynamics of businesses' interactions with their digital audience. At the forefront of this AI revolution are Large Language Models (LLMs) and Natural Language Processing (NLP) models, scientific breakthroughs that transcend the conventional boundaries of human-computer interaction.

LLMs, with their colossal neural networks and powerful language understanding capabilities, redefine the very essence of digital communication. Exemplified by pioneers like GPT-3, these models have the capacity to comprehend and generate human-like text, unlocking a new dimension of possibilities in e-commerce [1]. LLMs enable machines to not just respond but to converse, empathize, and understand the nuanced language of consumers.

NLP, a counterpart in the AI family, empowers machines to decode, interpret, and respond to human language in a manner that exceeds mere syntax [2]. This linguistic advancement allows e-commerce platforms to bridge the gap between user intent and machine understanding, facilitating a more intuitive and natural interaction between consumers and the digital storefront.

### **1.1.1. Enhancing the E-Commerce Experience through AI**

The integration of AI technologies, specifically LLMs and NLP models, in the e-commerce ecosystem marks the beginning of a revolutionary phase in transforming customer experiences. These technologies go beyond mere algorithms; they contribute to shaping a digital environment characterized by personalization, smooth operation, and responsiveness.

- **Improved Personalization:** AI, with its capacity to analyze extensive data sets and comprehend user behavior, empowers a degree of customization that surpasses the conventional boundaries of targeting based on demographics. Language models like LLMs and NLP models unravel the subtleties of individual preferences, crafting personalized shopping experiences to align with the distinct tastes and requirements of every individual consumer [3].
- **Streamlined Operations:** AI serves as a discreet conductor, optimizing processes for optimal efficiency and precision. Whether it's managing inventory or fulfilling orders, AI algorithms,

driven by data-driven insights, guarantee the seamless operation of every aspect of the e-commerce machinery. This not only automates processes but also contributes to enhanced customer satisfaction.

- **Data-Driven Decision-Making:** The union of AI and e-commerce stands as a testament to the prowess of data-centric decision-making. Language models like LLMs and NLP models navigate through vast datasets, extracting actionable insights that empower businesses to make well-informed decisions [4]. Whether it's suggesting products or devising pricing strategies, AI-driven analytics acts as the guiding compass steering e-commerce enterprises through the intricate landscapes of the digital marketplace.

### **1.1.2. Key Competitive Advantages Using AI**

In the relentless quest for a competitive edge, AI emerges as a powerful ally, equipping e-commerce platforms with capabilities that redefine industry standards.

- **Improved Customer Service through Chatbots:** Empowered by LLMs and NLP models, AI-powered chatbots redefine the concept of customer service. These virtual assistants provide instant, 24/7 support, addressing inquiries, resolving problems, and participating in natural, context-aware conversations [5]. The outcome goes beyond efficient customer service, creating an interactive and engaging digital storefront experience.
- **Enhanced Product Recommendations:** The age of generic recommendations fades away as AI-driven personalization takes center stage. Leveraging LLMs and NLP models, the recommender systems scrutinize user preferences, purchase history, and real-time behavior to provide tailored product recommendations that align with individual tastes [6]. This doesn't just enhance the shopping experience; it also fosters heightened customer satisfaction and loyalty [3].
- **Efficient Search and Discovery Mechanisms:** Leveraging the natural language understanding of LLMs, search algorithms are enhanced, allowing users to express their queries in a conversational manner [7]. NLP-driven search engines decode the context of user intent, providing results that are more accurate and relevant. This not only minimizes search friction but also empowers users to seamlessly explore and discover products.

The integration of AI in the e-commerce domain holds paramount significance, presenting robust tools to both customers and businesses. These tools not only result in time savings but also contribute to the enhancement of the overall user experience.

## ***1.2. The Imperative of Web Accessibility in E-Commerce***

As we navigate through the remarkable transformations brought about by AI in the e-commerce domain, it becomes crucial to address an aspect of digital commerce that is fundamental yet often overlooked: web accessibility. Accessibility in the e-commerce context transcends mere convenience—it embodies the principle of creating a digital marketplace that is accessible, ensuring that everyone, including people with disabilities, can navigate, understand, and interact with online platforms [8].

The significance of web accessibility is undeniable. It broadens the market reach by ensuring e-commerce platforms are universally accessible, embodying the principles of web accessibility [9]. By incorporating accessibility features, e-commerce entities not only break down barriers but also demonstrate their dedication to delivering an equitable shopping experience. This accessibility, consequently, widens the audience scope, allowing businesses to access a market segment that has traditionally been underserved.

### **1.2.1. Beyond Moral and Legal Obligations**

While the adherence to web accessibility standards is commonly discussed in the context of moral and legal compliance, the argument for improving accessibility holds significant business merit. In the digital era, consumers are attracted to brands that align with their values and ethics, and accessibility stands out as a crucial consideration. Improving web accessibility, therefore, transcends mere regulatory fulfillment; it represents a strategic investment in constructing a brand that connects with a broader demographic, encompassing the over one billion individuals worldwide living with various forms of disabilities [10].

The rationale extends beyond ethical considerations; it is firmly grounded in robust business logic. E-commerce enterprises, through expanding the accessibility of their platforms, tap into a consumer market segment that holds substantial size and purchasing power. Additionally, accessible websites generally provide an improved user experience for all customers, not solely those with disabilities, fostering higher satisfaction and loyalty [11]. Essentially, by giving precedence to web accessibility, e-commerces go beyond mere compliance with standards; they actively elevate their market appeal and competitive edge.

## ***1.3. Progressive Web Apps: Bridging Accessibility and Engagement in E-Commerce***

In the pursuit of establishing a digital marketplace that is both inclusive and accessible, Progressive Web Apps (PWAs) emerge as a technological paradigm reshaping the user experience online. The PWAs essentially represent websites utilizing cutting-edge web technologies to provide a user experience that is virtually identical to native mobile applications [12]. Crafted to leverage the complete capabilities of web browsers, they bring functionalities that were traditionally exclusive to native apps.

### **1.3.1. Defining Progressive Web Apps**

PWAs seamlessly integrate the expansive accessibility of the web with the immersive interaction found in mobile applications. Crafted using conventional web technologies such as HTML, CSS, and JavaScript, these applications prioritize harnessing cutting-edge browser features to provide users with a dynamic and app-centric user interface [13]. The distinctive feature of PWAs lies in their universal compatibility, extending their functionality to every user, irrespective of their chosen browser or device—a testament to the foundational principles of progressive enhancement [14].

### **1.3.2. Core Features of PWAs**

Several key features make PWAs a compelling choice for enhancing web accessibility and user experience in e-commerce platforms:

- **Offline Functionality:** PWAs employ service workers to pre-cache essential resources, enabling seamless loading and functionality even in the absence of an internet connection [8]. This offline capability guarantees users access to content and the ability to perform various actions, such as browsing products and reading product descriptions, regardless of their connectivity status.
- **Fast Loading Times:** User experience and engagement hinge significantly on speed, making it a pivotal factor. PWAs prioritize speed optimization, employing strategies like cache-first networking to facilitate swift loading of web pages, even in challenging network conditions [12]. This heightened responsiveness contributes to a seamless user experience, effectively minimizing bounce rates and fostering extended engagement times.
- **App-Like Experience:** PWAs deliver an immersive, full-screen encounter akin to native applications [15]. With functionalities like home screen icons, push notifications, and access to device hardware such as the camera and GPS, PWAs create a robust and engaging user interface, ensuring a compelling experience that encourages repeated user engagement.

### **1.3.3. Enhancing Engagement and Conversion Rates**

PWAs go beyond merely enhancing web accessibility; they stand as a potent tool for elevating engagement and conversion rates within the realm of e-commerce. The app-like experience they provide cultivates a feeling of familiarity and ease of use. When coupled with offline functionality and swift loading times, this combination significantly amplifies user satisfaction [12]. Satisfied users are inclined to engage more profoundly with the platform, delving into product exploration, dedicating more time to the site, and ultimately completing purchases.

Furthermore, PWAs possess the capability to send push notifications, empowering e-commerce platforms to sustain user engagement beyond their site visits [8]. These notifications serve as reminders for abandoned carts, unveil new product launches, or present personalized discounts. This direct communication channel with users proves invaluable in promoting recurrent visits and nurturing brand loyalty.

Enhancing the user experience through attributes like offline functionality, rapid loading times, and an app-like interface, PWAs not only enhance the accessibility of e-commerce platforms but also elevate their engagement levels [13]. This heightened engagement plays a pivotal role in driving increased conversion rates, as a smooth and enjoyable shopping experience directly impacts purchasing decisions.

## ***1.4. Search Engine Optimization (SEO) for Elevated E-Commerce Visibility and Growth***

In the ever-changing landscape of digital commerce, where the battle for consumer attention is intense and constantly evolving, Search Engine Optimization (SEO) stands out as a key element for e-commerce triumph. Going beyond a simple digital marketing technique, SEO embodies a holistic strategy that extends beyond a website, influencing its visibility, relevance, and accessibility to a broad online audience [16].

Fundamentally, SEO involves fine-tuning a website to secure a higher position in search engine results, guaranteeing enhanced visibility to prospective customers in search of pertinent products or information [17]. In the realm of e-commerce, where the virtual shelf space is expansive yet fiercely competitive, SEO emerges as an essential strategy for businesses aspiring to establish a unique and prominent presence in the digital marketplace.

### **1.4.1. The Relationship Between Rankings, Traffic, and Business Growth**

The core of the SEO story lies in the link between high search engine rankings and a boost in web traffic [18]. Acting as the guardians of the digital realm, search engines serve as the main gateway for users seeking products, services, or information. Websites that secure top positions in search results gain digital prominence, leading to increased visibility and, consequently, a significant increase in targeted traffic [17].

When strategically directed, this increased traffic becomes the driving force behind sustainable business growth. Each click symbolizes a potential customer, and every visit presents an opportunity for engagement and conversion [19]. The path from search engine rankings to heightened web traffic is essentially a transformative journey that enables e-commerce platforms to broaden their reach, establish brand authority, and nurture enduring customer relationships.

### **1.4.2. Using SEO for Shaping E-Commerce Success**

The SEO landscape is dynamic, adjusting to the constantly evolving algorithms of search engines and the shifting behaviors of online users [16]. To comprehend the modern SEO landscape and its implications for e-commerce, it is crucial to spotlight recent trends that have reshaped the dynamics of digital visibility:

- **Content Relevance:** In addition to keywords, modern SEO places significant emphasis on the relevance of content [20]. Search engines prioritize content that goes beyond being merely keyword-rich; it should also be contextually relevant, valuable, and aligned with user intent. For e-commerce platforms, this means creating product descriptions, blog posts, and landing pages that deeply resonate with the needs and interests of the target audience.
- **User Experience (UX):** The user experience has become a crucial element influencing SEO rankings [21]. Search engines prioritize websites that provide seamless, intuitive, and satisfying user experiences. E-commerce platforms need to fine-tune their site architecture, navigation, and overall design to guarantee a smooth journey for users from initial discovery to the final checkout.
- **Mobile Optimization:** As mobile device usage escalates; search engines give preference to mobile-friendly websites [22]. E-commerce platforms require responsive designs that adjust to different screen sizes, guaranteeing a uniform and user-friendly experience for customers who access the site from smartphones and tablets.
- **Technical SEO:** The technical foundations of a website are pivotal for achieving SEO success [20]. Elements such as site speed, crawlability, and structured data markup significantly influence a website's search engine performance. E-commerce platforms that invest in technical SEO initiatives not only boost their online visibility but also provide search engines with essential signals for optimal indexing.

In essence, staying abreast of these trends is not merely an option but a necessity for e-commerce success. A strategic alignment with the SEO principles positions e-commerce to not only navigate the current SEO landscape but also to thrive in the face of evolving algorithms and user expectations.

## ***1.5. Revolutionizing Mobile E-Commerce with Accelerated Mobile Pages (AMP)***

As the digital landscape pivots toward mobile-centricity, the imperative for a smooth, lightning-fast user experience on handheld devices grows. In answer to this demand, Accelerated Mobile Pages (AMP) emerges as a revolutionary technology, not just expediting the loading times of mobile web pages but also reshaping the entire mobile e-commerce experience.

### **1.5.1. AMP and Its Relevance to Mobile E-Commerce**

AMP stands as an open-source framework designed to create web pages with nearly instantaneous load times on mobile devices [23]. Its speed is attained by streamlining the HTML, CSS, and JavaScript components of a webpage, guaranteeing an optimized and efficient rendering process [24]. In the realm of mobile e-commerce, where attention spans are fleeting and seamless interactions are pivotal, AMP serves as a catalyst for elevating the overall user experience.

### **1.5.2. Benefits of AMP for E-Commerce Websites**

- **Faster Page Loading Times:** The main advantage of AMP lies in its capacity to significantly reduce page loading times [25]. Through the removal of non-essential elements and the utilization of a streamlined coding structure, AMP-enabled pages load almost instantaneously, guaranteeing that potential customers are not discouraged by slow loading times. This rapid responsiveness is especially vital for capturing the impulsive nature of mobile shoppers.
- **Improved Mobile Search Engine Rankings:** Search engines prioritize pages that are both mobile-friendly and load quickly in their rankings [22]. Incorporating AMP offers a competitive edge by aligning with search engine algorithms, resulting in enhanced visibility in mobile search results. E-commerce websites that leverage AMP are more likely to feature prominently, catching the attention of users actively searching for products or services on mobile devices.

- **Enhanced User Experience on Mobile Devices:** In addition to its speed benefits, AMP enhances the user experience on mobile devices [26]. The streamlined and standardized design of AMP pages guarantees a consistent and intuitive browsing experience across different devices and screen sizes [27]. This not only cultivates user satisfaction but also lowers bounce rates, enticing users to delve into and engage with the content.

### **1.5.3. Potential of AMP to Increase Visibility and Engagement in E-Commerce Sales Funnel**

- **Increased Visibility:** The rapid loading times and mobile-friendly features of AMP play a crucial role in boosting visibility, particularly in mobile search results [22]. E-commerce platforms that embrace AMP are well-placed to seize the attention of users during the crucial searching phase, elevating brand visibility and enticing potential customers.
- **Higher Engagement Rates:** AMP facilitates a smooth and responsive browsing experience, allowing users to navigate through product listings, view images, and initiate transactions with ease [27]. This heightened engagement not only enhances the overall user experience but also boosts the likelihood of conversions.
- **Supporting the E-Commerce Sales Funnel:** AMP aligns with the foundational principles of a well-structured e-commerce sales funnel. It guarantees swift-loading pages during the discovery/search phase, facilitates seamless navigation through product pages, and streamlines the checkout process, making AMP an integral element in guiding users through the sales funnel [28]. The outcome is a user-friendly journey that transitions seamlessly from discovery to conversion.

Accelerated Mobile Pages emerge as a game-changer in the mobile e-commerce landscape. The fusion of speed, mobile optimization, and an elevated user experience positions AMP as a pivotal technology for e-commerce websites, aiming not only to meet but surpass the expectations of mobile consumers.

## ***1.6. Collective Contribution: Necessity for E-commerce Growth and Sustainability***

The integration of these technologies, when viewed collectively, transcends the realm of trend and manifests as a necessity for e-commerce businesses aspiring for growth and sustainability in the digital age. AI augments the customer experience, PWAs bridge accessibility and engagement, SEO establishes visibility and authority, and AMP accelerates mobile experiences. The synergy of these innovations creates an ecosystem where user expectations are not just met but exceeded, forming the foundations for long-term customer relationships and business success.

In this thesis, the overall objective is not only to discern all the advancements and technologies that could enhance e-commerce sustainability but also to apply each of these advancements to real-world and practical scenarios, extracting valuable outcomes regarding their suitability, efficiency, and the benefits derived for enhancing e-commerce competitiveness.

In **Chapter 2**, an in-depth exploration of LLMs and NLP is conducted. The research delves into the complex world of NLP, where computer science, linguistics, and AI converge. A comprehensive analysis of the Generative Pre-trained Transformer (GPT) architecture follows, shedding light on its complexities. The examination then transitions to the processes of Pre-Training and Supervised Fine-Tuning, designed specifically to meet the unique requirements of LLMs. Expanding on this, the discourse covers the art and science of prompting engineering and provides a concise overview of the open-source LLaMA-2 model, highlighting its sophisticated architecture.

The research extends to evaluating LLMs' effectiveness in analyzing sentiment within e-commerce product reviews. A comparison is made between two cutting-edge LLMs (GPT-3.5 and LLaMA-2) and two renowned NLP models (BERT and RoBERTa). The evaluation assesses their ability

to perform sentiment analysis on product reviews, examining their capabilities both before and after being fine-tuned for this specific purpose. The study aims to discern the potential of LLMs in enhancing customer satisfaction analysis, making a significant contribution to the e-commerce domain, with promising findings indicating high predictive accuracy rates.

An innovative approach is also introduced for enhancing product recommender systems through the integration of unsupervised learning models—specifically, K-means clustering, content-based filtering (CBF), and hierarchical clustering—with the advanced capabilities of the GPT-4 LLM. The core innovation of this methodology lies in the application of GPT-4 for the evaluation phase, leveraging its sophisticated natural language understanding to significantly improve the precision and relevance of product recommendations. Experimental results underscore the effectiveness of this integrated framework, demonstrating its superiority in refining recommender systems. This represents a noteworthy advancement in the field of recommender systems, providing e-commerce businesses with an efficient and scalable tool to improve their product recommendation capabilities.

Addressing the growing importance of sustainability in various industries and the increasing consumer demand for environmentally friendly products a GPT-powered software is introduced for automating the identification of sustainable features in product descriptions. Utilizing the GPT-3.5, the software is capable of classifying products into distinct sustainability categories and extracting crucial indicators of sustainability, such as the use of eco-friendly materials, energy efficiency, recyclability, and ethical sourcing practices. The automation facilitated by this software obviates the necessity for manual analysis, thereby streamlining the process of evaluating product sustainability. This technology not only enables consumers to make more informed and sustainable purchasing decisions but also assists businesses in highlighting their commitment to environmental stewardship. Experimental results underscore the efficacy and precision of the software in pinpointing sustainable product features, attesting to the GPT model's potential in the domain of sustainability assessment.

Finally, **Chapter 2** highlights the strategic importance of mitigating customer churn in the dynamic e-commerce environment, where substantial customer acquisition costs necessitate a cost-effective approach to retaining customers. It underscores the significance of predictive models in identifying potential churners not only for increasing customer lifetime value but also for enhancing brand reputation and sustaining profitability. To achieve these objectives, the research employs two well-established models trained on a specialized dataset tailored for churn modeling. The primary goal is to assess the precision levels of these models in predicting customer churn. The results indicate that both models demonstrate high accuracy, with the neural network trained model slightly outperforming the logistic regression trained model in terms of precision. The research further highlights the critical role of dataset quality and balance in the training process, emphasizing their impact on the models' predictive capabilities. This strategic approach aims to guide businesses towards enduring success by addressing customer churn and fostering long-term customer satisfaction.

**Chapter 3** delves into the dynamic landscape of PWAs and their intersection with web accessibility, a crucial aspect in today's digital realm where an ever-expanding user base interacts with websites, e-commerce platforms, and mobile applications on a daily basis. This exploration unfolds against the backdrop of the significant rise in mobile app usage, highlighting the importance placed on their usability and user experience (UX). This is further accentuated by the emergence of hybrid apps seamlessly integrating web technologies with native functionalities. Despite their advantages, both hybrid and native apps face challenges, such as the need for ample device storage and complexities associated with cross-platform upgrades.

In response to these limitations, Google introduced PWAs in 2015, aiming to blend native app features within mobile browsers through modern application programming interfaces (APIs). PWAs promise enhanced capabilities, reliability, and installability across devices using a unified codebase. The study recognizes a concerning trend: as web technologies progress, attention towards accessibility for individuals with disabilities appears to diminish, despite the growing importance of inclusive access to new technologies. This issue is exacerbated by the lack of accessibility in academic and government websites, further complicating the challenges faced by people with disabilities.

Through a comprehensive literature review on PWAs and accessibility technologies, driven by the urgent need for accessibility solutions across various platforms, the research critically examines PWAs, covering their technological benefits and limitations. Additionally, it provides a critical analysis of current accessibility technologies, with a specific focus on the recently introduced Web Content Accessibility Guidelines (WCAG) 2.2 draft.

The primary objective of this research is to compare and evaluate PWA as a technology against other (non-PWA) web technologies in terms of their compatibility with existing accessibility guidelines. This evaluation aims to determine whether PWAs offer superior accessibility compared to their counterparts.

By synthesizing insights on PWAs and accessibility technologies, the study aspires to promote enhanced web accessibility for all users, advocating for a more accessible digital world. The challenges and opportunities discussed at the convergence of PWAs and web accessibility in this chapter lay a solid foundation for future research. The multifaceted contribution of this research not only broadens the understanding of PWAs and their potential for enhancing web accessibility but also underscores the necessity for ongoing advancements in technology that are inclusive and accessible to everyone, regardless of their abilities.

**Chapter 4** provides a comprehensive exploration of SEO, covering both on-page and off-page SEO techniques. Beyond merely focusing on SEO Techniques, it suggests leveraging existing web technologies to boost SEO, thoroughly examining each technology's potential to enhance SEO strategies. The study also examines black hat SEO techniques and paid traffic.

Special attention is drawn to the increasing reliance of websites on digital marketing, particularly SEO, for sustainable growth in today's dynamic digital landscape. The integral link between website traffic and developmental success intensifies the engagement of SEO experts who utilize premium SEO audit tools. These tools scrutinize a website's source code, pinpointing structural changes essential to meet specific ranking criteria (SEO factors). However, the considerable costs associated with these tools and specialists pose financial barriers.

In response to these challenges, the research introduces an open-source Python-based Machine Learning SEO software tool. This tool offers a cost-effective alternative for both website owners and SEO specialists, conducting meticulous analyses of top-ranking websites for designated search terms and providing tailored recommendations to optimize website performance, enabling them to surpass competitors.

Through illuminating case studies, particularly one centered on a competitive keyword, the software's efficacy in significantly boosting organic traffic for an e-commerce platform is vividly showcased. When applied in a practical e-commerce context, the tool delivers substantial results, offering guidance that elevates average daily organic traffic from 10 to 143 visitors.

Additionally, this chapter addresses pivotal questions regarding crucial SEO techniques, strategies employed by renowned websites and e-commerce platforms, and the determination of the priority for implementing these SEO techniques. A specialized PHP-based SEO tool is developed to crawl websites and analyze applied SEO techniques, generating a comprehensive database detailing the SEO techniques employed by each entity.

The research synthesizes these findings into a figure illustrating the usage score of various SEO techniques among the top-listed websites and e-commerce platforms. This figure functions as a practical guide-map for Webmasters. By aligning their strategies with those implemented by highly-ranked websites, even those lacking specialized SEO knowledge can effectively apply the recommended techniques. The overarching objective is to assist Webmasters in elevating their website's Search Engine Ranking Positions (SERPs) and attracting more traffic by adopting proven and effective SEO practices.

Further exploration delves into the impact of SEO strategies on airline and hotel websites during the COVID-19 era, leading to the adoption of new digital marketing strategies to navigate heightened competition. Both industries have made substantial investments in optimizing their websites to enhance rankings on SERPs, resulting in increased organic traffic and conversions. These

studies aim to reveal the specific SEO strategies employed by airline and hotel websites, encompassing their e-commerce booking platforms, during this challenging period. To achieve this, an internally developed tool is employed to scan the source code of each website, extracting information about the implemented SEO techniques. Concurrently, four third-party APIs are utilized to gather real-time data, including domain authority, organic keywords, backlinks, and web traffic over a 12-month period. Through meticulous data analysis, the study identifies the most widely adopted SEO techniques and explores their impact on web traffic for airline and hotel websites. The research establishes a significant correlation between SEO techniques and metrics, web traffic, and subsequently, increased conversions and bookings for airlines and hotels. By isolating the most prevalent SEO techniques, the studies suggest that these strategies, when applied by large companies like airlines and hotels, could also yield positive results for smaller websites.

Lastly, the SEO techniques and technologies discussed earlier are applied to a live E-commerce platform to assess and validate their impact on organic traffic. The results of this implementation reveal a remarkable increase of 782% in organic traffic. Additionally, the E-commerce site achieves first-page rankings for multiple keywords in search results, indicating the tangible and positive impact of the implemented SEO techniques.

**Chapter 5**, investigates the exponential growth of mobile subscriptions, which surpassed 7.7 billion by late 2017, and addresses the persistent issue of suboptimal user experiences for mobile users. Despite the widespread use of mobile devices, challenges such as slow-loading web pages and the lack of responsive design technologies persist. In response to discussions with news publishers and technology companies aimed at improving mobile web performance, Google introduced the AMP project in October 2015.

The AMP project represents an open-source HTML framework crafted to simplify the creation of rapid, seamlessly loading web pages prioritizing user experience. The study explores AMP Usage and Restrictions, as well as the application of AMP technologies to a real-life E-commerce project, presenting a comprehensive comparative analysis between Organic and AMP traffic.

The **Conclusion** chapter provides a brief overview of the meticulously conducted research, shedding light on its innovative contributions within the dynamic e-commerce landscape. The integration of AI, NLP, LLMs, Web Accessibility, PWAs, SEO, and AMPs stands as an indispensable framework for e-commerce enterprises aspiring to thrive and endure in the ever-evolving digital era.

The synergistic incorporation of these cutting-edge technologies forms an ecosystem where user expectations transcend mere fulfillment to reach levels of surpassing satisfaction. This cohesive integration not only establishes a solid foundation for enduring customer relationships but also serves as a catalyst for overall e-commerce success. The collective impact of AI, NLP, LLM, Web Accessibility, PWAs, SEO, and AMPs goes beyond individual enhancements, creating a holistic and innovative approach that aligns seamlessly with the evolving demands of the contemporary e-commerce landscape.

## 2. Large Language Models (LLMs) and Natural Language Processing (NLP) in E-commerce

*Adapted from:*

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Precision-Driven Product Recommendations Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems. *Software* 2024. <https://doi.org/10.3390/software3010004>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Natural Language Processing Journal* 2024, 6, 100056. <https://doi.org/10.1016/j.nlp.2024.100056>

Roumeliotis, K.I.; Tselikas, N.D. Churn Modeling in E-commerce: Leveraging the Power of Logistic Regression and Neural Networks. *Conference on Business Intelligence and Modelling, Springer Proceedings in Business and Economics*. (Acceptance Letter. Pending DOI)

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Unveiling Sustainability in Ecommerce: GPT-Powered Software for Identifying Sustainable Product Features. *Sustainability* 2023, 15, 12015. <https://doi.org/10.3390/su151512015>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. *Preprints* 2023, 2023072142. <https://doi.org/10.20944/preprints202307.2142.v2>

Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* 2023, 15, 192. <https://doi.org/10.3390/fi15060192>

### 2.1. Introduction

Large Language Models (LLMs) have revolutionized the Artificial Intelligence (AI) landscape through their advanced Natural Language Processing (NLP) capabilities. Characterized by massive neural networks with billions of parameters, these models excel in understanding and generating human-like text [29]. In the realm of E-commerce, LLMs have played a crucial role in enhancing customer interactions through chatbots, virtual assistants, and personalized recommendation systems. These models leverage techniques such as transfer learning, enabling them to generalize from a broad range of data sources and adapt to specific E-commerce domains [2].

By employing sentiment analysis, entity recognition, and language understanding algorithms, LLMs can interpret user queries, discern customer preferences, and provide tailored product suggestions [1]. This not only streamlines the shopping experience but also contributes to increased customer engagement and satisfaction. Moreover, the continuous training and fine-tuning of LLMs with evolving datasets ensure that these models stay abreast of market trends and user behaviors, making them indispensable tools for the dynamic landscape of E-commerce.

The subsequent sections of **Chapter 2** are structured as a cohesive exploration of various aspects related to NLP and LLMs. **Section 2.2** delves into the multifaceted realm of NLP, situated at the intersection of computer science, linguistics, and AI. In **Section 2.3**, the focus shifts to a detailed analysis of the architecture characterizing the Generative Pre-trained Transformer (GPT). Moving forward, **Section 2.4** intricately examines the nuanced Pre-Training Phase and Supervised Fine-Tuning processes, specifically tailored to address the demands of LLMs. **Section 2.5** extends this inquiry to encompass an examination of the prompting engineering. **Section 2.6** provides an overview of the

opensource LLaMA-2 model. Subsequently, Sections 2.7 to 2.10 present our published articles separately, highlighting the application of AI models in E-commerce.

## ***2.2. Natural Language Processing (NLP)***

NLP holds immense significance in the field of AI, as it enables machines to comprehend, interpret, and manipulate human language. NLP plays a crucial role in bridging the gap between human communication and machine understanding [30]. It encompasses a wide range of tasks, including speech recognition, text classification, sentiment analysis, machine translation, information extraction, and question answering [30]. NLP techniques allow for the extraction of meaningful insights from vast amounts of unstructured textual data, facilitating efficient information retrieval, analysis, and decision-making processes [31]. NLP has numerous practical applications, such as chatbots, virtual assistants, language translation services, content summarization, and sentiment analysis in social media. By unlocking the power of language, NLP enables the development of intelligent systems that can understand and communicate with humans in a more natural and intuitive manner.

## ***2.3. Generative Pre-trained Transformer (GPT) Architecture***

In recent years, the significance of language models, particularly the groundbreaking GPT developed by OpenAI, has witnessed an exponential surge across a broad spectrum of applications [2]. The GPT model has exhibited exceptional aptitude in producing coherent and contextually pertinent textual content [2].

The introduction of transformer architecture has brought about a revolution in the fields of NLP and machine translation [32]. This breakthrough was enabled by the self-attention mechanism, which facilitates parallel processing and enhances the capture of long-range dependencies [2]. The incorporation of self-attention layers in the transformer allows the model to weigh the significance of individual words during prediction, leading to improved context comprehension [32]. By replacing traditional recurrent neural network (RNN) or convolutional neural network (CNN) processing, the transformer achieves faster training and inference times [33]. As a foundational framework, the transformer has elevated language-understanding tasks, serving as the basis for state-of-the-art NLP models such as BERT, GPT, and RoBERTa [2]. Its success emphasizes the critical role of attention mechanisms in deep learning and stimulates continuous research in the field of NLP. The transformer's ability to process sequences in parallel contributes to its scalability and adaptability to parallel hardware systems, enhancing its widespread utilization across diverse applications and architectures [34]. Moreover, the encoder-decoder structure within the transformer significantly improves machine translation and language generation, representing a pivotal advancement in the realm of NLP [35].

## ***2.4. Pre-Training Phase and Supervised Fine-Tuning***

The pre-training phase of GPT (Generative Pre-trained Transformer) constitutes a pivotal step wherein the model is immersed in vast and diverse text data originating from various sources [2]. By harnessing the self-attention mechanism inherent in the transformer architecture, GPT gains the ability to comprehend the contextual relationships within the text and engage in unsupervised learning to predict subsequent words in a sentence [35]. GPT models lack true comprehension and knowledge of the world; they do not possess genuine understanding of syntax, semantics, or discourse in the way humans do. Instead, they rely on statistical associations and patterns they have learned from the vast amounts of text during training [36]. While they can generate impressive responses, especially in context, they are fundamentally different from human intelligence. As a result of the pre-training phase, GPT acquires the adaptability to be fine-tuned for specific tasks, granting it versatility in generating

contextually appropriate text. The exposure to diverse textual data further enriches GPT's understanding of language and enhances its recognition of contextual nuances [37,38]. The resounding success of the pre-training phase can be attributed to GPT's remarkable capacity to assimilate knowledge from diverse sources and process vast volumes of data, laying the foundational groundwork for its exceptional language-understanding capabilities [2].

The fine-tuning phase of GPT, a critical step in its development and is essential for optimizing performance in specific downstream tasks. During this phase, the pre-trained model undergoes further training using task-specific data that include labeled examples [2]. By fine-tuning, GPT can adapt its language representation to meet the unique nuances and requirements of the target task [30]. This process involves adjusting the model's weights and biases to achieve optimal performance on the specific task at hand [39]. The fine-tuning process empowers GPT to hone its language generation capabilities, ensuring contextually appropriate responses for the given task [5]. Moreover, it facilitates GPT's specialization in a diverse array of NLP tasks, encompassing text classification, summarization, and sentiment analysis [5]. Through fine-tuning on diverse datasets, GPT becomes amenable to customization for various real-world applications, elevating its practicality and versatility [2]. The success of the fine-tuning phase is intricately tied to the quality and size of the task-specific data used for training, as these factors significantly impact the model's performance on the targeted task [40,41].

## ***2.5. Prompts in GPT Models***

In the context of GPT models, prompts refer to specific instructions or starting points provided to guide the generation of text [42]. Prompts serve as input cues that help shape the output generated by the model, allowing users to influence the direction and content of the generated text [43]. These prompts can take various forms, such as sentences, questions, or even keywords, and are designed to elicit desired responses from the model [43]. The use of prompts provides a means for users to convey their intentions and preferences, effectively acting as a way to interact and collaborate with the language model [42].

Prompts play a crucial role in guiding the generation of text in GPT models. When a prompt is provided as input, the model leverages its pre-trained knowledge to understand the context and generate text that is coherent and relevant to the given prompt [43]. The initial prompt acts as a starting point, influencing the subsequent text-generation process. The model's understanding of grammar, vocabulary, and contextual dependencies allows it to build upon the prompt and generate coherent and contextually appropriate text. By adjusting the prompt, users can control the tone, style, or specific content of the generated text, enabling a more personalized and tailored output [2,44]. This flexibility in utilizing prompts empowers users to shape the output according to their desired outcomes, making GPT models a versatile tool for various applications.

Constructing effective prompts requires careful consideration to achieve desired outcomes. One approach is to provide specific instructions or constraints within the prompt to guide the model's response. For example, specifying a format, desired length, or specific keywords can influence the generated text accordingly [44]. Another technique is to incorporate context within the prompt by including relevant background information or setting the stage for a particular topic [45]. This helps the model generate text that aligns with the intended context. Additionally, iterative prompting can be employed, where users can progressively refine or add to the initial prompt to shape the ongoing text generation process [45]. Experimentation and fine-tuning of prompts are essential to optimize the model's performance and ensure the desired output [44]. By carefully crafting prompts, users can effectively guide GPT models and harness their capabilities to generate text that meets their specific requirements or objectives [42].

## 2.6. LLaMA-2 model

The LLaMA and LLaMA-2 models are instances of GPT models, built upon the original Transformers architecture [46]. The LLaMA models employ GPT-3-like pre-normalization, utilizing the `RMSNorm` normalizing function at the input of each transformer sub-layer [47]. This approach enhances training stability by rescaling the invariance property and implicit learning rate adaptation ability. Additionally, LLaMA benefits from the `SwiGLU` activation function, replacing the conventional `ReLU` non-linearity activation function, leading to improved training performance [46].

Incorporating insights from the GPT-Neo-X project, LLaMA incorporates rotary positional embeddings (RoPE) at each layer, contributing to its overall performance [46]. Notably, LLaMA-2 introduces essential architectural differences which include an increased context length, doubling the context window size from 2048 to 4096 tokens [47]. This extension enables the model to handle more extensive information, proving beneficial for tasks involving long documents, chat histories, and summarization.

Furthermore, LLaMA-2 implements a grouped-query attention (GQA) format with eight key-value projections, addressing the complexity concerns associated with the original Multi-Head attention baseline [46]. This modification proves effective in managing the increased context windows or batch sizes.

As a result of these updates, LLaMA demonstrates significantly improved performance across various tasks, surpassing or closely matching other specialized GPT models such as Falcon and MPT [48]. The model's promising performance paves the way for further research, anticipating future comparisons with prominent closed-source models like GPT-4 and Bard.

### 2.6.1. LLaMA-2 Models and Versions

Meta has developed and publicly released the LLaMA-2 family of LLMs, comprising a set of pre-trained and fine-tuned generative text models with parameter sizes spanning from 7 billion to 70 billion [46]. The parameter size plays a crucial role in determining model accuracy, with larger parameter sizes signifying extensive training with vast amount of data, leading to more precise and dependable responses. Among these, the fine-tuned LLMs, specifically named LLaMA-2-Chat, have been tailored to optimize performance in dialogue-based use cases. Through comprehensive benchmark assessments, LLaMA-2-Chat models have demonstrated superior capabilities compared to open-source chat models. Additionally, in human evaluations focused on assessing helpfulness and safety, LLaMA-2-Chat models have shown comparable performance to some prominent closed-source models, including ChatGPT and PaLM [49,50].

### 2.6.2. Pretraining Phase

The training corpus of LLaMA-2 comprises a novel blend of publicly accessible data sources, excluding any data originating from Meta's products or services. During the data selection process, diligent measures were taken to exclude data from websites known to contain substantial volumes of personal information about private individuals. The model underwent training on an extensive dataset comprising 2 trillion tokens, exhibiting twice the context length of its predecessor, LLaMA-1 [51]. This design choice strikes a balance between performance and computational cost, with a deliberate emphasis on up-sampling the most factual sources to enhance knowledge while mitigating potential hallucination issues [46].

The developers of LLaMA-2 retained much of the pretraining settings and model architecture employed in LLaMA-1. The model adheres to the standard transformer architecture proposed by Vaswani [32], utilizing pre-normalization with `RMSNorm` [52] and the `SwiGLU` activation function [53]. Furthermore, it integrates rotary positional embeddings (RoPE) [46,54].

Key differences between LLaMA-1 and LLaMA-2 lie in the augmentation of context length and the adoption of grouped-query attention (GQA). These architectural modifications contribute to the improved capabilities of LLaMA-2 and its ability to handle more extensive contextual information during language generation tasks [46].

### 2.6.3. LLaMA-2 Fine-tuning

LLaMA-2 is pre-trained using publicly available online data [50]. An initial version of LLaMA-2-chat is then created through the use of supervised fine-tuning. Next, LLaMA-2-chat is iteratively refined using Reinforcement Learning from Human Feedback (RLHF), which includes rejection sampling and proximal policy optimization (PPO) [46].

In the pursuit of optimizing the performance of LLMs for dialogue-style instructions, the key aspects revolve around the quality and diversity of third-party Source-Free Tuning (SFT) data [54]. Although numerous sources provide such data, their limited diversity and quality led the focus to prioritize the collection of high-quality SFT examples, resulting in significant improvement. Meta's study also found that a limited set of clean instruction-tuning data could yield satisfactory outcomes, and approximately tens of thousands of SFT annotations were sufficient for achieving high-quality results [46]. Notably, the annotation platform and vendor choices influenced the downstream model performance, emphasizing the significance of thorough data checks. The validation process confirmed the high quality of outputs from the SFT model, suggesting a potential to reallocate annotation efforts towards preference-based annotation for RLHF. The investigation encompassed 27,540 annotations, excluding Meta user data, and drew parallels with related research highlighting the effectiveness of focusing on quality over quantity in instruction-tuning endeavors [46].

The RLHF is a model training approach used to further align the behavior of a fine-tuned language model with human preferences and instructions [55]. Human preference data is collected, where annotators select their preferred choice between two model outputs, aiding in the training of a reward model that automates preference decisions [56]. The collection procedure involves a binary comparison protocol to maximize prompt diversity, with annotators rating the degree of preference for their chosen response. Safety and helpfulness aspects are specifically focused on, allowing the application of distinct guidelines to each. The reward modeling data, referred to as "Meta reward modeling data," comprises over one million binary comparisons, surpassing existing open-source datasets in terms of conversation turns and average length [46]. Continuous updates to the reward model are essential to adapt to the evolving LLaMA-2-Chat iterations and maintain accurate rewards for the latest model.

The reward model plays a crucial role in RLHF, where it takes a model response and its corresponding prompt, and outputs a scalar score indicating the quality in terms of helpfulness and safety [56]. By leveraging these response scores as rewards, the RLHF process optimizes LLaMA-2-Chat to align better with human preferences and enhance helpfulness and safety [46]. To address the trade-off between helpfulness and safety, two separate reward models are trained - one for each aspect. The reward models are initialized from pretrained chat model checkpoints to ensure knowledge transfer and prevent information mismatch. Training objectives involve converting collected human preference data into binary ranking labels, with a margin component to handle different preference ratings [6,46]. Additionally, the reward models are trained on a combination of newly collected data and existing open-source datasets to improve generalization and prevent reward hacking.

## 2.7. *LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation*

### 2.7.1. Section Summary and Significance of the Study

In today's rapidly evolving digital landscape, the e-commerce sector has emerged as a significant economic force, with revenue projected to surge to an impressive US\$4.18 trillion by 2024, signaling its crucial role in shaping the future of global commerce [57]. It is essential to emphasize that customer satisfaction remains and will continue to remain at the forefront of any domain associated with e-commerce [58]. Nevertheless, comprehending the customer's emotional state and satisfaction following a purchase requires considerable human resources and ongoing advanced data analysis.

One approach to gauge customer satisfaction following a purchase is the evaluation and comprehension of each product review [59]. Understanding the underlying significance of each product review holds paramount importance for e-commerce platforms, marketplaces, and product manufacturers [60]. Questions like the customer's contentment, encountered issues during shipping or product utilization are among the inquiries e-commerce owners and manufacturers seek answers to, in order to optimize their services. However, analyzing each review individually, considering the multitude of product reviews associated with each product in an e-commerce platform, is a time-consuming and economically impractical endeavor [61].

The solution to the aforementioned issue is addressed by data analysts employing advanced statistical and Machine Learning (ML) models that systematically extract information from each review. Subsequently, this information is conveyed to the relevant team for the purpose of adjusting the e-commerce strategy. Some of the ML models previously utilized for related tasks included Transfer Learning Models [62], Naive Bayes [63], Logistic Regression [64], Semantic Analysis [65], and others. However, with the advancement of AI, the advent of deep learning, and the revolution introduced to the world in early 2023 by LLMs, many tasks previously performed with older models can now be executed much more efficiently through the use of LLMs.

This research aims to evaluate the effectiveness of two prominent LLMs in comprehending and extracting information from product reviews. The LLMs utilized in this study are the GPT-3.5 by OpenAI [66] and the open-source LLaMA-2 by Meta [46]. Initially, the two base models will be assessed based on their performance in predicting product review ratings. Subsequently, these models will undergo fine-tuning using a large dataset of product reviews, and their performance will be assessed following the fine-tuning process. Concurrently, fine-tuning will be applied to two NLP models, BERT and RoBERTa, using the same training set. Comparisons will be made both within the NLP models and between the NLP models and the LLMs.

The primary aim of this research is twofold: firstly, to assess and compare the performance of the models, and secondly, to address specific research questions that have not been adequately answered by prior studies:

- Q1: Which LLM demonstrates superior efficacy in assessing product reviews?
- Q2: Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?
- Q3: How significant is the process of fine-tuning LLMs for domain-specific tasks?
- Q4: What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?
- Q5: To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?
- Q6: What impact does fine-tuning have on the generated output of LLMs?
- Q7: Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?
- Q8: Can LLMs be effectively utilized for the evaluation of product reviews, and how can LLMs revolutionize the e-commerce sector?

To address the aforementioned research questions, the research initiates by providing a concise literature review in **Section 2.7.2**. This review aims to assist readers in gaining a deeper understanding of topics related to customer sentiment analysis and satisfaction, as well as the utilization of NLP and LLMs for this particular study. **Section 2.7.3** outlines the methodology employed in facilitating the current research, encompassing the data collection of product reviews, dataset cleaning, the creation of a universal prompt, model execution, fine-tuning, and the utilization of these models to identify the lexical components that influenced both reviewers and LLMs in assigning ratings to product reviews. **Section 2.7.4** presents the research findings, while **Section 2.7.5** not only addresses the research questions, but also compares pre- and post-fine-tuning models. Key insights from the authors'

observations are also highlighted in this section. Finally, the research concludes in **Section 2.7.6**, by discussing implications and providing directions for future research work.

## **2.7.2. Literature review**

### **Customer satisfaction in E-commerce**

In the realm of e-commerce, the foundation of success is customer satisfaction [67]. In the constantly shifting world of online retail, where countless choices are merely a click away, contented customers play a pivotal role in ensuring the endurance and prosperity of e-commerce enterprises. Furthermore, customer satisfaction (CS) is closely linked with customer retention (CR) and loyalty (CL) [68]. Happy customers are not only more inclined to return for future purchases, but also to become loyal advocates, creating a ripple effect by recommending the e-commerce platform to their network of friends and family [69]. Significantly, retaining an existing customer often costs less than acquiring a new one [70]. Consequently, customer satisfaction emerges as an economically efficient strategy for sustained business growth [71].

In the interconnected digital realm, satisfied customers transform into influential brand ambassadors. They eagerly share their positive experiences through social media, review platforms, and word-of-mouth endorsements. These spontaneous recommendations can significantly enhance an e-commerce business's reputation and visibility, thereby attracting new customers and broadening its market reach [72].

Furthermore, customer satisfaction exhibits a direct connection with sales and revenue. Pleased customers tend to spend more, make repeat purchases, and show a higher lifetime customer value [71]. Additionally, favorable reviews and ratings from satisfied customers actively stimulate sales by providing social proof and enhancing the perceived trustworthiness of the e-commerce platform [73].

High customer satisfaction levels, therefore, act as a safeguard against customer attrition. E-commerce businesses operate in a fiercely competitive landscape, where dissatisfied customers or those experiencing subpar service are prone to switch to a competitor [67]. Consequently, maintaining high levels of customer satisfaction becomes crucial not only to prevent customer loss, but also to ensure long-term business sustainability.

It is evident that customer satisfaction can serve as a crucial differentiator. E-commerce businesses that consistently provide exceptional customer experiences gain a competitive advantage, which, in turn, draws and retains customers [74]. This competitive edge can lead to increased market share and, ultimately, enhanced profitability.

The feedback obtained from contented customers represents a valuable resource for e-commerce enterprises looking to innovate and enhance their offerings. Customer satisfaction surveys, reviews, and direct feedback offer essential insights into what is working well and what needs improvement [75]. This information empowers e-commerce businesses to adapt and evolve in response to shifting customer preferences and market dynamics.

### **Product reviews in understanding customer sentiment**

In today's consumer landscape, product evaluations have evolved into a pivotal element of decision-making processes. The rise of e-commerce and digital marketplaces has forged a deep reliance on reviews among customers seeking insights into a product or service's performance, quality, and overall satisfaction [76]. These reviews serve as a wellspring of knowledge, guiding consumers toward well-informed purchase choices, molding perceptions of products and services, and holding businesses accountable for their offerings [77].

The digital realm, enriched by online platforms, social media, and e-commerce websites, has opened up a vast arena for individuals to share their encounters and viewpoints on various products and services. This profusion of product appraisals has revolutionized how consumers approach their buying decisions. A survey from Askalidis et al. (2017) underscores this shift, revealing that a staggering 95% of shoppers peruse online reviews before committing to a purchase, and the purchase

likelihood for a product with five reviews is 270% greater than the purchase likelihood of a product with no reviews, which strongly underscores the surging significance of product evaluations in the consumer decision-making process [78].

Diverse sources, including e-commerce websites, social media platforms, specialized review sites, and mobile applications, serve as reservoirs of these product assessments. In these platforms, customers thoughtfully recount their experiences, meticulously outlining both the favorable and unfavorable aspects of a given product or service. These appraisals take on a variety of forms, from textual narratives to numerical ratings, star-based evaluations, or a fusion of these elements [79].

Understanding customer sentiment from these reviews requires the application of sophisticated NLP and sentiment analysis methodologies. Sentiment analysis algorithms scrutinize the language, tone, and contextual nuances within the reviews, categorizing them as either positive, negative, or neutral [80]. These algorithms leverage ML and linguistic analysis to extract valuable insights from textual data.

Product evaluations have a multifaceted impact on consumer behavior [81]. First and foremost, they furnish invaluable insights into a product's performance, features, and quality. These reviews bridge the information gap between customers and products, empowering consumers to make more judicious choices [82]. Additionally, these reviews play a crucial role in building trust. Numerous positive evaluations for a product can establish a sense of credibility and reliability, while a lack of reviews or an excess of negative ones may trigger caution for potential buyers [78].

Product reviews serve not only consumers but also provide value to manufacturers and retailers. They act as feedback mechanisms, offering valuable insights into areas for improvement and gauging customer satisfaction. Companies can leverage this feedback to enhance their products and services, address customer concerns and suggestions, and adjust their marketing strategies accordingly [60].

### **LLMs in E-commerce**

Despite being in an early stage compared to other NLP models, recent research has demonstrated the efficiency of LLMs in various tasks related to e-commerce. Kanaan et al. (2023) introduce a novel approach using GPT for product category matching, showcasing the superior effectiveness of the GPT-based model [83]. Chen et al. (2023) explore LLMs for relation labeling in e-commerce Knowledge Graphs, demonstrating remarkable learning capabilities and the potential to replace human labeling [84].

Gao et al. address limitations of LLMs and Specialized Translation Models (STMs) in e-commerce translation, introducing an LLMs-based E-commerce machine translation approach (LEMT) that outperforms state-of-the-art Neural Machine Translation (NMT) models [85]. Wang and Na (2023) propose a novel approach to enhance e-commerce search and recommendation systems by converting structured data into textual data, allowing for more effective search and recommendation using LLMs [86].

Al Wahshat et al. (2023) focus on combating manipulated reviews in e-commerce platforms using GPT-4, demonstrating its efficacy in identifying and flagging manipulated reviews with remarkable accuracy [87]. Zhou et al. (2023) automate product description generation in e-commerce using the LLaMA-2 7B language model, emphasizing the potential of LLMs in optimizing e-commerce [88].

Continuing the research, Roumeliotis et al. (2023) introduce a software powered by GPT to automatically identify sustainable product features in e-commerce, aiding informed, sustainable purchasing decisions [29]. Shi et al. (2023) address limitations of general LLMs in handling personalized features unique to e-commerce, introducing LLaMA-E, achieving state-of-the-art results in e-commerce authoring tasks [89]. Liu et al. (2023) explore the synergy between conversational recommender systems (CRSs) and LLMs, highlighting their complementary strengths in improving pre-sales dialogue tasks [7].

Ma et al. (2023) tackle challenges in applying LLMs to specific domains by focusing on domain-specific continual pre-training, demonstrating its effectiveness in improving few-shot learning and zero-shot performance in the e-commerce domain [90]. Maragheh et al. (2023) investigate using item aspects generated by LLMs to improve ranking tasks, introducing Theme-Aware Keyword Extraction (LLM TAKE) for enhanced keyword generation in e-commerce [91].

Finally, Orzoł and Szopik-Depczyńska (2023) explore how ChatGPT can enhance customer communication and boost sales in the e-commerce industry, providing insights into the potential benefits of ChatGPT for e-commerce stores [92]. Li et al. (2023) contribute by addressing challenges of applying general LLMs to e-commerce tasks, proposing the first E-commerce instruction dataset, EcomInstruct, and introducing EcomGPT, a specialized model demonstrating superior zero-shot generalization capabilities [93].

Collectively, these studies emphasize the substantial potential of LLMs in revolutionizing and enhancing various facets of the e-commerce landscape.

### **2.7.3. Research methodology**

This study aims to unravel the intricate emotional nuances embedded within product reviews through the proficient application of LLMs, with a particular focus on the widely recognized GPT-3.5 model and the novel open-source LLaMA-2 model. Our primary objective is to gain a profound understanding of the psychological states that drive reviewers when composing their assessments, whether they manifest satisfaction, dissatisfaction, or a complex ambivalence toward the products they have purchased.

To evaluate how proficiently LLMs can determine the sentiments expressed in individual words within product reviews, we began our study by using these models to predict star ratings for a large dataset of product reviews. We hypothesized that since LLMs can efficiently predict the rating stars of a review, they also have the capability to understand the meaning and emotions conveyed by the words in each review.

To assess the effectiveness of LLMs for this specific task without bias, we concurrently employed two well-known NLP models (non-LLM): BERT and RoBERTa. These models were tasked with making predictions for the same tasks as the LLMs, generating valuable results for meaningful comparisons.

In the preliminary phases of our research, we meticulously scrutinize the ability of LLMs to accurately discern the intricate emotional subtleties expressed by product reviewers, conducting a comprehensive performance comparison between the two aforementioned models. Following this, we undertake the task of fine-tuning both LLMs and their non-LLM counterparts for sentiment analysis within the context of product reviews. This fine-tuning process enhances their capacity to respond with increased precision to the multifaceted emotional dimensions present in reviews.

Confirming through our research that fine-tuned LLMs exhibit significantly higher performance compared to their base counterparts, we advanced our investigation one step further by deploying fine-tuned models to identify specific lexical elements — words within product reviews that influenced their selection of review rating stars. We compiled these words into a table, which will aid future studies in comprehending how LLMs perceive the sentiment of a reviewer by analyzing the words they choose to employ in their product reviews.

The entire codebase utilized for the research, spanning from the Flask-based Chrome application designed for review collection to the fine-tuning process of the models, along with the associated datasets, is available in a GitHub repository with an MIT open source license [94].

To accomplish the objectives of this research, given the volume and complexity of the processes involved, it became imperative to adopt a specific research methodology. This methodology has been carefully selected to facilitate both the development of the software for review collection and the subsequent execution and fine-tuning of the models. The detailed framework for these procedures is presented in the following subsections to provide a comprehensive and coherent strategy for the study.

## **Chrome app development and data collection**

The research was initiated with the development of a Chrome application, meticulously engineered using object-oriented Python, Flask, and JavaScript. This application was purposefully designed to facilitate the systematic extraction of product reviews from prominent online marketplaces, including industry giants like Amazon and eBay.

The Chrome application was crafted with precision to ensure its reliability and effectiveness in data retrieval. Leveraging object-oriented Python, a structured and modular codebase was established, allowing for efficient maintenance and future enhancements. The utilization of Flask, a Python web framework, streamlined the development process, facilitating seamless interactions between the frontend and backend components. Additionally, JavaScript played a pivotal role in collecting and transmitting data to the Flask server.

The primary objective of the application was to collect comprehensive information from online product listings. The researchers systematically gathered a multitude of data points, including product titles, product descriptions, product categories, and user-generated product reviews. The product titles and descriptions provided valuable insights into the attributes and features of the items under scrutiny. The product categories enabled efficient product classification and categorization. The focus, however, was on user reviews, extracting the review title, the complete body of the review, and the accompanying star rating given by the reviewer.

The acquired data, extensive and diverse, was systematically stored within an SQLite database. This relational database management system ensured data integrity and offered a structured environment for organizing and querying the collected information. Each data point, from product descriptions to user reviews, was indexed, allowing for efficient retrieval and analysis.

The extensive scope of the data collection efforts resulted in a substantial repository of information. The application successfully collected data from 616 distinct products, creating a robust dataset consisting of 5,029 product reviews.

Beyond being the foundation for the research project, the data collected through the Chrome application holds significance in both academic and practical contexts. This dataset provides numerous opportunities for subsequent analysis, including sentiment analysis, trend identification in product reviews, and the examination of potential correlations between product attributes and review ratings. Moreover, it provides valuable insights for businesses, enabling a deeper understanding of customer satisfaction and preferences in the highly competitive e-commerce landscape.

## **Dataset splitting and preprocessing**

To ensure the quality and effectiveness of our predictive modeling and fine-tuning processes, we prepared the dataset, adhering to a systematic approach. These processes involved several key steps, each designed to enhance the suitability of the data for subsequent analysis.

Data preprocessing plays a pivotal role in enhancing the quality of the dataset. In this stage, we addressed several aspects of text data to make it more amenable for modeling. One of the primary tasks in data preprocessing involved the removal of special characters and extra white spaces, critical for maintaining data cleanliness and consistency. Special characters, if left unattended, can lead to noise in the dataset and hinder the effectiveness of NLP techniques.

Text normalization is also essential for ensuring that text data is consistent and standardized. As part of this process, we performed operations such as converting accented characters to their base form. This step is especially important for languages with diacritics, as it ensures that words with accents' variations are treated uniformly.

Furthermore, to ensure that text data is case-insensitive, we uniformly converted all text to lowercase. This avoids the distinction between, for example, "Review" and "review", ensuring that NLP models are not sensitive to the case of text.

In addition, in an effort to make the most of the textual information within the reviews, we merged the review titles and bodies. This consolidation step simplifies the input data and ensures that the complete content of each review is considered when making predictions.

In order to gauge the number of tokens within each review, essential for processing by a LLM, we employed the `tiktoken` Python library. This provided us with a close estimation of the number of tokens, ensuring that we stay within the capacity limits of the LLM while maintaining the integrity and completeness of the text.

In the realm of NLP tasks, splitting a dataset is pivotal for effective model development, refinement, and assessment. This practice plays a crucial role in achieving robust and reliable models capable of generalizing adeptly to new, unseen data across diverse linguistic contexts. In our specific approach, we partitioned our dataset into training (70%), validation (15%), and test (15%) sets, employing the `train_test_split` function from `scikit-learn` (training: 3520, validation: 754, test: 755) [95]. This partitioning occurs in two stages: an initial split generates a training set and a temporary set, with the latter subsequently undergoing further division into validation and test sets. The training set serves a critical role during the fine-tuning phase, where the model acquires patterns, relationships, and representations from the input data, enabling it to make predictions or execute specific tasks. The validation set is crucial for the model tuning phase, involving adjustments to hyperparameters and configuration. Evaluating the model's performance on the validation set informs modifications that enhance generalization and curbed overfitting. Finally, the fine-tuned models employ the test set for making predictions, completing the comprehensive process of model fine-tuning and evaluation.

To highlight the importance of the dataset during the fine-tuning of NLP models, subsequent to splitting the dataset containing 5,029 product reviews, we created a smaller dataset comprising approximately half of the training and validation data (training: 1,757, validation: 365). This reduced dataset was then utilized to re-fine-tune the same base models, making predictions on the same test set (test: 755).

### Creating a compelling prompt understandable to both LLMs

By employing well-established prompting engineering techniques, our goal was to create a prompt that is compatible with various LLMs, enhancing the accessibility of their output through our code. Our approach prioritizes not only the content of the prompt but also the formatting of the output.

Creating a universal prompt starts with a deep understanding of the diverse LLM landscape. Each model, such as GPT-3, GPT-4, LLaMA-2, and others, have unique characteristics, strengths, and limitations. Designing a prompt that can elicit meaningful responses from these models while maintaining accessibility is a complex challenge. To address this, we applied two well-established prompting engineering techniques, considering the idiosyncrasies of various LLMs [96].

1. **Model-Agnostic Content:** Our approach involves developing a prompt that is model-agnostic, meaning it does not rely on the specific architecture or knowledge of any particular LLM. This enables our prompt to be adaptable and transferable across a range of models. We focus on creating a prompt that convey the task clearly, providing relevant context and information that any LLM can comprehend.
2. **Accessibility through Output Formatting:** An often-overlooked aspect of prompt design is the format of the output. We strongly emphasize designing a code-friendly output format that enhances accessibility. This is achieved by structuring the responses in a coherent and intuitive manner. For the specific requirements of this task, it was necessary that the outputs be structured in JSON format.

After numerous iterations and experiments with both LLMs, the final prompt, comprehensible to both models, effectively elicited responses in the desired output format, is illustrated in **Figure 1**.

```
prompt = 'Assign integer star ratings (between 1 and 5) to the following product reviews.' \
        'Return your response in JSON format like this example:' \
        '{"rating1":integer, "rating2":integer, ...}.' \
        'Do not provide explanations or justifications for the ratings. Reviews"\n'
```

Figure 1. Model-Agnostic prompt

### **Model deployment, fine-tuning LLMs and predictive testing**

In this study, four NLP models were employed: the BERT model, the RoBERTa model, and the LLMs GPT-3.5 and LLaMA-2. For each of these models, a specific methodology was followed for both fine-tuning and product review rating prediction.

#### *GPT-3.5 model*

In this phase, the `gpt-3.5-turbo-1106` base model was initially employed to make predictions for the review ratings in the test set using a specific prompt presented in **Figure 2**. Leveraging their extensive training, LLMs have acquired the ability to discern subtle cues influencing reviewer sentiment, enabling them to generate predicted star ratings. Our methodology involved inputting the language model with the textual content of reviews, allowing it to comprehend nuances, sentiments, and underlying themes in both review titles and their bodies. To make predictions and perform fine-tuning on the `gpt-3.5-turbo-1106` base model, we utilized the official OpenAI API. Additionally, based on statements from Azure CTO Mark Russinovich [97], Azure employs Low-Rank Adaptation (LoRA) Parameter-Efficient Fine-Tuning (PEFT) and DeepSpeed techniques to reduce GPU usage and enhance memory efficiency during the fine-tuning of its GPT-3 model.

The predicted ratings generated by GPT-3.5 were then meticulously compared with the original ratings provided by the reviewers. This comprehensive analysis allowed us to evaluate the effectiveness of the base model in capturing the essence of reviews and predicting ratings that align closely with human judgments.

During the fine-tuning phase, the `gpt-3.5-turbo-1106` model was exposed to extensive data and fine-tuned to adapt to specific nuances and patterns in the training set. The multi-epoch training approach allowed the model to iteratively refine its understanding and capabilities, progressively enhancing its performance on target tasks and ensuring readiness for accurate predictions and valuable insights in subsequent evaluations.

Two fine-tuning processes were conducted, one identified by the Job ID `ftjob-fduMpiYlj23jZ17b9wXooyBd` and the other (reduced dataset) by `ftjob-P6xOmuaYW4JiIxxTkCg6ytpF`. The former processed 1,468,176 tokens with training loss at 0.0574 and validation loss at 0.0216 across three epochs. The latter processed 730,491 tokens with a training loss of 0.3690 and a slightly higher validation loss of 0.3927 in three epochs. These metrics indicate the models' adaptability and generalization during fine-tuning.

After the completion of the fine-tuning phase, the fine-tuned models were assigned the task of predicting star ratings for product reviews in the same test set. The obtained results were then integrated into an `SQLite` table, simplifying further comparative analysis.

To facilitate the fine-tuning of the LLMs, two `JsonL` files were created, encompassing prompt and completion pairs, as illustrated in **Figure 2**.

```

# Llama 2 JsonL Sample
{"prompt": "You are a product reviewer. Assign integer star ratings (between 1 and 5)"
  "to the following product reviews."
  "Return your response in JSON format like this example:"
  "{\"rating1\":integer, \"rating2\":integer, ...}."
  "Do not provide explanations or justifications for the ratings. Reviews\\n\\n"
  "1. almost perfect. these would be perfect if they had pockets",
"completion": "{\"rating1\":4}"}

# GPT 3.5 JsonL Sample
{"messages": [{"role": "system", "content": "You are a product reviewer"},
  {"role": "user", "content":
    "Assign integer star ratings (between 1 and 5)to the following product reviews."
    "Return your response in json format like this example"
    "{\"rating1':integer, 'rating2':integer,...}."
    "Please avoid providing additional explanations. Reviews:\\n\\n"
    "1. almost perfect. these would be perfect if they had pockets"},
  {"role": "assistant", "content": "{\"rating1\\":4}"}]}

```

Figure 2. Fine-tuning JSONL samples.

### *LLaMA-2 model*

In this phase, the llama-2-70b-chat base model was initially used to predict review ratings in the test set using a specific prompt from **Figure 2**. Predicted ratings from llama-2-70b-chat were carefully compared with the original ratings provided by reviewers. Given the significant computational resources required by LLaMA models, similar to GPT-3.5, we leveraged the Replicate API for predictions and fine-tuning. Similar to Azure’s use of LoRA PEFT techniques to reduce GPU requirements during the fine-tuning of its GPT-3 model, Replicate employs LoRA – specifically, QLora techniques – for the fine-tuning the LLaMA model.

Two fine-tuning processes were performed, identified by the Job IDs ecommerce-reviews5029:4a107315 [98] and (reduced dataset) ecommerce-reviews50:b05681bd [99]. The first job utilized 8x A40 (Large) GPUs, with a total runtime of 377.02 min. Training lasted 22,394.4514 s, achieving a speed of 0.465 samples per second, 0.116 steps per second, and a final training loss of 0.462 over three epochs. In the second job, the fine-tuning process also utilized 8x A40 (Large) GPUs, with a total runtime of 182.11 min. Training lasted 10,708.3497 s, achieving a speed of 0.478 samples per second, 0.12 steps per second, and a final training loss of 0.4918 over three epochs.

After the completion of fine-tuning, the fine-tuned models predicted star ratings for product reviews in the test data, and the results were also integrated into an SQLite table for further comparative analysis.

### *BERT model*

In this phase, we utilized the bert-base-uncased variant [100] for our NLP task. Specifically, we employed the BertForSequenceClassification model from the transformers library. The foundational BERT model captures contextualized representations of input tokens through its transformer layers. BertForSequenceClassification, a distinct iteration of BERT, incorporates an additional classification head pre-configured for sequence classification tasks. Typically, this classification head consists of a fully connected layer that transforms the BERT output into class probabilities [101].

The bert-base-uncased model is a case-insensitive iteration of BERT, wherein all input text is converted to lowercase during training. Similar to the broader BERT architecture, this variant is transformer-based, featuring multiple layers and hidden units. The architectural specifications of the bert-base-uncased model include 12 layers, 768 hidden units, 12 heads, and 110 million parameters [102]. The self-attention mechanisms integrated into BERT facilitate the capture of contextual dependencies within input sequences.

During the fine-tuning process, the `bert-base-uncased` model was employed for sentiment classification using the training set comprising reviews and their corresponding ratings. The fine-tuning and prediction processes were executed in Google Colab, utilizing an A100 GPU. The dataset, loaded from a CSV file in Google Drive, underwent preprocessing, which included tokenization using the BERT tokenizer [103]. Hyperparameters for the fine-tuning process were specified, including a learning rate of  $2e-5$ , a batch size of 8, and utilizing the Adaptive Moment Estimation (Adam), Adam with Weight Decay (AdamW) and Stochastic Gradient Descent (SGD) optimizers. The model underwent training for three epochs, with progress tracked using the `tqdm` library [104]. The training loop encompassed backpropagation, optimization, and validation on a separate dataset. The code checked for GPU availability for faster computation. Following training, the fine-tuned model and tokenizer were saved to a directory in Google Drive for future use and predictions. Our code embodies a systematic and thorough fine-tuning methodology, assuring the model's alignment with the unique requirements of the sentiment classification task and its preparedness for deployment. After the fine-tuning process, the model proceeded to generate predictions for the test sample. The corresponding code, the training and validation loss, and the validation accuracy for this process are available in an `ipynb` Jupyter file on GitHub [94].

### *RoBERTa model*

In this phase, we employed the `roberta-base` model, accessible at [105], for sentiment classification. Following the BERT-base architecture, the RoBERTa model architecture incorporates numerous layers of self-attention mechanisms, facilitating the effective capture of contextual relationships within input sequences. The specific configuration of the RoBERTa model involves a 12-layer transformer-based neural network with 768 hidden units, 12 attention heads, and a total of 125 million parameters [102]. Each transformer layer within this architecture typically integrates self-attention mechanisms and feedforward neural networks, collectively enhancing the model's proficiency in comprehending and representing intricate relationships within sequential data. For the prediction phase, the `RobertaForSequenceClassification` model was utilized from the `transformers` library [106], involving the loading of the pre-trained model through the `pretrained` method.

The same fine-tuning methodology like the BERT model followed for the RoBERTa model, tailored for sentiment classification using a custom dataset of reviews and corresponding ratings. The fine-tuning and prediction processes were also executed in Google Colab, utilizing an A100 GPU. The training set, sourced from a CSV file in Google Drive, undergoes preprocessing, employing the RoBERTa tokenizer to tokenize text data and preparing it as `PyTorch` tensors. The hyperparameters governing the fine-tuning process, including a learning rate of  $2e-5$ , a batch size of 8, and the AdamW, Adam, and SGD optimizers, are specified to optimize model parameters effectively. The training unfolds over three epochs, with GPU utilization for accelerated computation when available. The training loop, visualized using the `tqdm` library, encompasses both training and validation phases, evaluating the model's accuracy. Post-fine-tuning, the RoBERTa model and tokenizer are saved for future use in a designated Google Drive directory. Subsequently, the model engaged in the generation of predictions for the test set. The script and associated resources, such as the training and validation loss, as well as the validation accuracy, are available on GitHub in a Jupyter notebook file (`ipynb`) for broader accessibility and collaboration [94].

### **Linguistic factors influencing product review ratings**

In the final phase of the research, the fine-tuned models trained on the entire training set were selected, along with those reviews to which the LLMs assigned a rating equal to that provided by the human reviewer. The fine-tuned LLMs were then prompted to identify specific lexical components, specifically the words contained within product reviews that significantly influenced the assignment of rating stars. For both models, a universal prompt was used, as presented in **Figure 3**, and their responses were saved in a CSV file to facilitate analysis.

```

prompt = f'''Please carefully analyze the review text provided and identify the words or phrases
that played a significant role in influencing both your and the reviewer's rating decision.
Your response should be in JSON format, listing the words or phrases
and their corresponding influence on the rating.
For example: {"1": "word or phrase", "2": "word or phrase", "3": "word or phrase", ...} \n
Review text: {review_text}\n
Review rating: {review_rating}'''

```

**Figure 3.** Model-agnostic prompt for identifying lexical elements in product reviews

This analysis aimed to uncover and clarify how specific words or phrases significantly influenced the overall assignment of rating stars by reviewers. The ultimate goal was to enhance our understanding of the linguistic factors that impact the determination of product review ratings, thus improving our insight into the decision-making processes employed by both human reviewers and fine-tuned LLMs.

#### 2.7.4. Comparative analysis and models’ evaluation

In **Section 2.7.3**, we elaborate on the methodological framework employed to evaluate the predictive capabilities of four NLP models, including GPT-3.5 and LLaMA-2 LLMs, BERT, and RoBERTa, in the context of star-rating product reviews and understanding their underlying nuances. This section presents the results of the comparative analysis conducted among the four models at various stages of our research.

##### Fine-tuned models evaluation

Before presenting our findings, it is crucial to emphasize the importance of model evaluation. Within the domains of ML and NLP, the process of evaluating models assumes a central role, affording us invaluable insights into the performance and effectiveness of our fine-tuned models. Model evaluation acts as a guiding compass, enabling us to make well-informed decisions regarding their suitability and, in parallel, propelling us toward progressive advancements in fine-tuning and optimization tailored to specific applications. **Table 1** presents a comprehensive range of evaluations for each model, encompassing a set of critical evaluation metrics.

**Table 1.** Model Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1 Score
base:gpt-3.5-turbo-1106	0.5483	0.553	0.5483	0.5467
base:llama-2-70b-chat	0.5086	0.5029	0.5086	0.5038
ft:gpt-3.5-turbo-1106 (100%)	0.6424	0.6416	0.6424	0.6409
ft:llama-2-70b-chat (100%)	0.6185	0.6162	0.6185	0.6147
ft:bert-adam (100%)	0.5881	0.5881	0.5881	0.5871
ft:bert-adamw (100%)	0.5841	0.5921	0.5841	0.5766
ft:bert-sgd (100%)	0.2225	0.0706	0.2225	0.0888
ft:roberta-adam (100%)	0.6066	0.6157	0.6066	0.6017
ft:roberta-adamw (100%)	0.5536	0.5951	0.5536	0.5325
ft:roberta-sgd (100%)	0.2225	0.0495	0.2225	0.081
ft:gpt-3.5-turbo-1106 (50%)	0.6212	0.6207	0.6212	0.6197
ft:llama-2-70b-chat (50%)	0.604	0.6009	0.604	0.5997
ft:bert-adam (50%)	0.5523	0.5575	0.5523	0.5532
ft:bert-adamw (50%)	0.5775	0.5741	0.5775	0.5736
ft:bert-sgd (50%)	0.1629	0.0269	0.1629	0.0462
ft:roberta-adam (50%)	0.5152	0.5305	0.5152	0.5181
ft:roberta-adamw (50%)	0.5497	0.5351	0.5497	0.5286
ft:roberta-sgd (50%)	0.1642	0.027	0.1642	0.0463

### **Base models evaluation phase**

In the initial phase of our research, we engaged both LLM base models, namely `gpt-3.5-turbo-1106` and `llama-2-70b-chat`, in the task of predicting the star ratings associated with the test set comprising 755 product reviews. Our primary objective was to assess the accuracy of these LLMs in predicting the star ratings originally assigned by users when crafting their reviews.

Upon careful examination of the responses generated by the base models and comparing them to the original star ratings provided by users, we made several notable observations. The `gpt-3.5-turbo-1106` model exhibited a satisfactory grade of accuracy, precisely predicting 54.83% of the star ratings, which is translated to a successful prediction in 414 out of the 755 reviews that constituted our test dataset. In contrast, the `llama-2-70b-chat` model, while still displaying noteworthy performance, achieved a slightly lower accuracy rate of 50.86%, i.e., accurate predictions for 384 out of the 755 reviews in our test set. These findings suggest that both LLMs possess satisfactory predictive capabilities, with the GPT-3.5 base model showing a marginally higher accuracy in this specific task.

It is essential to emphasize that achieving a 50% accuracy rate in predictive modeling is akin to making predictions equivalent to random chance. However, when dealing with a classification task involving more than two choices, such as in the case of review ratings with 5 options, the baseline accuracy due to random chance is calculated to be 20% (1 out of 5).

In practical terms, this implies that a predictive model that simply guesses without any learning or information should, on average, be correct 20% of the time. Therefore, models that surpass this baseline are providing more meaningful predictions. For instance, the LLM base models achieved prediction accuracies of 54.83% and 50.86%, respectively, demonstrating a considerable improvement over random chance in discerning the nuances of the given reviews.

These accuracy percentages signify that the models are making predictions significantly above what one would expect from random guessing, indicating their ability to capture patterns and relationships within the data, thereby enhancing the quality of predictions.

However, the efficiency of the models can be further enhanced by fine-tuning them for the specific task, improving their ability to capture and understand intricate patterns and relationships in the data.

### **Fine-tuned LLMs evaluation phase**

In the subsequent phase of our study, we conducted fine-tuning on the GPT-3.5 and LLaMA-2 models, simultaneously applying the same fine-tuning process to the BERT and RoBERTa models using three different optimizers. This process was carried out on the training set, consisting of 3,520 product reviews, with the objective of improving the models' performance and enhancing their adaptation to the task at hand.

Subsequently, the fine-tuned models were put to the test once more, tasked with the challenge of predicting star ratings for the same test dataset. The results of this phase were quite effective. The LLaMA-2 model, post fine-tuning, managed to correctly predict star ratings in a 61.85% of cases, i.e., 467 out of the 755 reviews included in the test set, showcasing a considerable improvement in its predictive capabilities. In contrast, the GPT-3.5 model demonstrated even more noteworthy progress following the fine-tuning process. It achieved a remarkable 64.24% accuracy rate, i.e., accurately predicting star ratings in 485 out of the 755 reviews within the test set. This substantial improvement in predictive accuracy highlights the effectiveness of fine-tuning in enhancing the performance of both LLMs, with GPT-3.5 exhibiting a particularly strong response to this optimization process.

At the same time, the non-LLM models responded to fine-tuning to a similar extent, with the BERT model achieving 58.81% correct predictions and RoBERTa reaching 60.66% using the ADAM optimizer. With the ADAMW optimizer, both models showed a slightly lower performance on predictions, i.e., 58.41% and 55.36%, respectively. In contrast, with the SGD optimizer, the results were highly disappointing at 22.25% for both models, indicating that this particular optimizer is not suitable for the given task.

### Evaluation of fine-tuned LLMs for half of the dataset

To assess the impact of the amount of training data in the fine-tuning process of an NLP model, we employed the same fine-tuning procedure on four models. However, this time, we utilized approximately 50% of the training data, specifically 1,757 samples out of the previously employed 3,520.

In this specific phase, the GPT-3.5 fine-tuned model continued to maintain its edge with a 62.12% accuracy rate in predicting star ratings, correctly predicting 469 out of the 755 test samples. Meanwhile, the LLaMA-2 model achieved a 60.4% accuracy, correctly predicting star ratings in 456 out of the 755 test samples.

In a corresponding manner, NLP models (non-LLM), fine-tuned on approximately half of the dataset, yielded BERT predictions of 52.23% and RoBERTa predictions of 51.52%, employing the ADAM optimizer. Employing the ADAMW optimizer resulted in predictions of 57.75% and 54.97%, whereas utilization of the SGD optimizer led to predictions of 16.29% and 16.42%.

### Assessing models' performance and proximity with original ratings

In order to gain meaningful insights into the performance of base and fine-tuned models and the degree to which their predictions align with the original ratings provided by reviewers, we utilized the `mean_absolute_error` class from the `sklearn` library. The Mean Absolute Error (MAE) was used to calculate the difference between the predictions of the base and fine-tuned models and the original ratings provided by the reviewers. The results are presented in **Figure 4**, **Figure 5** in the form of scatterplots. In these diagrams, the perfect fit line reflects the points where predictions are in perfect alignment with the actual values.

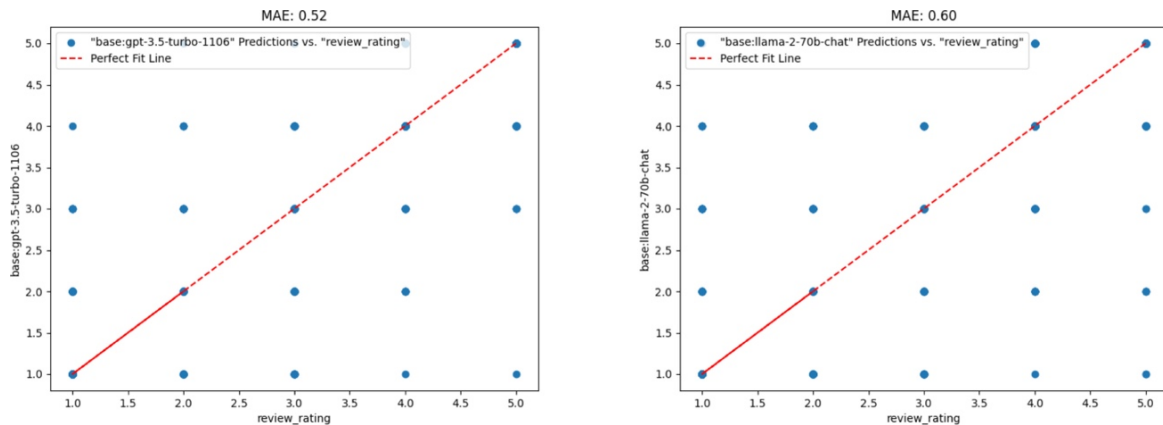


Figure 4. MAE for Base Models

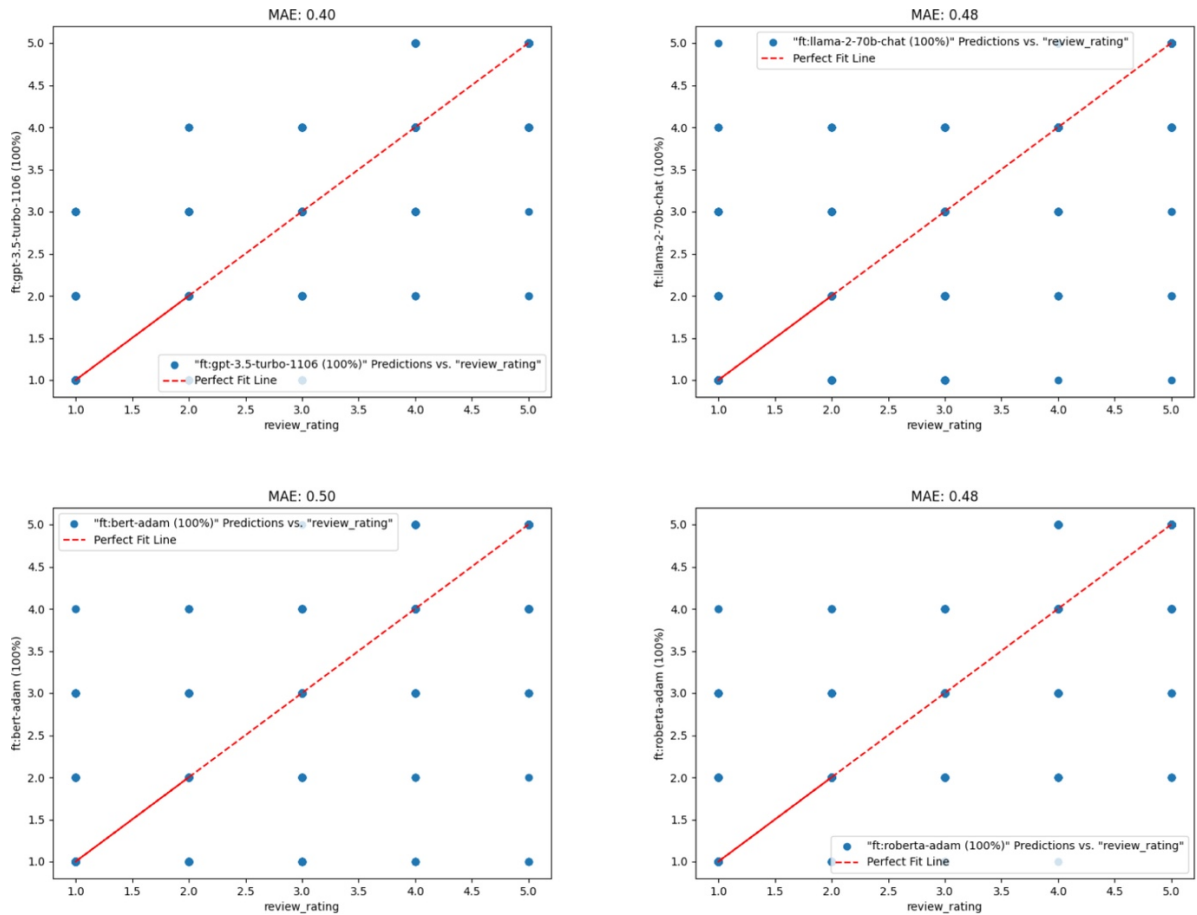


Figure 5. MAE for Fine-Tuned Models

The MAEs provide insights into the accuracy of various NLP models across different fine-tuning configurations. A lower MAE suggests that, on average, the model’s predictions are closer to the true values, indicating better accuracy, predictive capabilities, and overall performance. The base `gpt-3.5-turbo-1106` and base `llama-2-70b-chat` models exhibit MAEs of 0.5166 and 0.6026, respectively, indicating the average absolute differences between their predictions and true values. When fine-tuned on the full training set (100%), both GPT and LLaMA models show improvements, achieving much lower MAEs of 0.3974 and 0.4755, respectively.

Furthermore, the fine-tuned versions of BERT and RoBERTa on the same training set, with the ADAM optimizer providing better predictions according to our statistic results, yield MAEs of 0.5046 and 0.4741, respectively. Notably, in these evaluations, the RoBERTa model outperforms the BERT model.

### Evaluation of emotional nuances and lexical elements

In accordance with the methodology and logic presented in **Linguistic factors influencing product review ratings** section, the fine-tuned LLMs were called upon to distinguish those words and phrases within the product reviews that led them to correctly predict the reviewers’ ratings. For the purposes of the research, a sample of 5 reviews was chosen in which the predictions of both fine-tuned models matched the ratings given by the reviewers. The results were highly interesting and are presented in **Table 2**.

Table 2. Lexical elements detected by fine-tuned LLMs in review bodies

Review Content	Rating	ft_llama_100	ft_gpt_100
----------------	--------	--------------	------------

junk sound cuts inout for a while at first but following not long after it completely stopped working	1	['junk', 'sound', 'cuts inout', 'not long after', 'completely stopped working']	['junk', 'sound cuts inout', 'completely stopped working']
nice hiking shoes it fits well i returned it because i felt the sole wasnt padded enough and the bottom was a bit stiff	4	['nice hiking shoes', 'it fits well', 'returned', 'wasnt padded enough', 'stiff']	['nice', 'fits well', 'sole wasnt padded enough', 'bottom was a bit stiff']
this is great theres just one thing great shower curtain thats nice and light and soft the one thing i wish it had was magnets in the corners but other than that its great	4	['great', 'light', 'soft', 'magnets', 'corner']	['great', 'shower curtain', 'nice', 'light', 'soft', 'wish it had magnets', 'other than that']
soft good quality it is very soft and would recommend only flaw is the threading on the pillow cases are a little loose im sure not all manufactured are like that	4	['soft', 'good quality', 'very soft', 'recommend', 'loose', 'threading', 'pillow cases', 'manufactured', 'flaw']	['soft', 'good quality', 'recommend', 'threading on the pillow cases are a little loose']
nice nice color and stretchyfits a little baggy for slim fit	4	['nice', 'color', 'stretchy', 'fits a little baggy', 'slim fit']	['nice color', 'stretchy', 'fits a little baggy', 'slim fit']
nice shirt for the price looks nice pressed out very neat plenty long to stay tucked in	5	['nice shirt', 'for the price', 'looks nice', 'pressed out', 'very neat', 'plenty long', 'stay tucked in']	['nice', 'shirt', 'price', 'looks nice', 'pressed out', 'neat', 'plenty long', 'stay tucked in']
works good sound sucked but not the cord retured the xlr and mic and cant return this for some reason	3	['works good', 'sound sucked', 'not the cord', 'retured the xlr and mic', "can't return this"]	['good', 'sucked', 'cord', 'returned', 'xlr', 'mic', 'cant return']

In the majority of the reviews presented in **Table 2**, the words identified by the fine-tuned LLMs were the same. At the same time, when viewed from a human perspective, these words are indeed the most representative and justify the reviewers' ratings, highlighting the potential of LLMs to identify the most important words in a text in a similar way to humans. Undoubtedly, making a sweeping generalization would lack objectivity. However, judging from human evaluation, it appears that LLMs can understand the context behind lexical elements to a lesser but encouraging extent, similar to humans.

### 2.7.5. Research findings

In the preceding sections, we presented, the methodology and results for evaluating the predictive capabilities of GPT-3.5 and LLaMA-2 LLMs, as well as BERT and RoBERTa models in tasks related to the star-rating assessment of e-commerce product reviews. In this section, the research findings and the observations made by the authors that have emerged from the study concerning the utilization of LLMs for sentiment analysis on product reviews are being presented, while also addressing the research questions.

#### Base models comparison

- **Research Question 1:** Which LLM demonstrates superior efficacy in assessing product reviews?
- **Research Statement 1:** The `gpt-3.5-turbo-1106` base model outperforms the `llama-2-70b-chat` model by 3.97% in tasks associated with predicting star ratings for product reviews.

In the **Base models evaluation phase**, the research results regarding the performance of the base models `gpt-3.5-turbo-1106` and `llama-2-70b-chat` in predicting star-ratings in a test-sample of 755 product reviews were presented. Notably, it was observed that GPT-3.5 is significantly more accurate than the LLaMA-2 model. This particular difference is attributed to the training phase, where the GPT-3.5 model is assumed to have been trained with over 175 billion parameters [107], while the LLaMA-2 model has 70 billion parameters [46].

One plausible hypothesis to explain this deviation in performance lies in the distinct training processes that these two models underwent, particularly in their exposure to various public datasets. It is conceivable that GPT-3.5 benefited from a more extensive and diverse range of training data, including potentially more meticulous and specific fine-tuning. This expanded review training dataset might have provided GPT-3.5 with a deeper understanding of nuanced topics, enabling it to generate more accurate and contextually relevant responses.

In essence, the notable performance of GPT-3.5 underscores the importance of not only the quantity of parameters, but also the quality and diversity of the training data in achieving optimal language model capabilities.

#### **Fine-tuned models comparison**

- **Research Question 2:** Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?
- **Research Statement 2:** Following the fine-tuning process, the GPT-3.5 model exhibits superior performance to the LLaMA-2 model by a margin of 2.39% in tasks related to predicting star ratings for product reviews.

Based on the results obtained in **Fine-tuned LLMs evaluation phase**, the fine-tuned GPT-3.5 model demonstrates an accurate assessment of 64.24% of review ratings in comparison to the LLaMA-2 fine-tuned model, which predicts ratings with a success rate of 61.95%. Although the difference between the two models is not greater than 2.39%, the superiority of the fine-tuned GPT-3.5 model in comparison to the LLaMA-2 fine-tuned model, can be ascribed to several factors inherent in its architecture. Notably, the GPT-3.5 model showcases a more effective utilization of the training data during the fine-tuning process. This effectiveness can be attributed to several key characteristics of the GPT-3.5 architecture.

- First and foremost, the GPT-3.5 model benefits from a more extensive and diverse pre-training corpus, which provides it with a broader foundation of linguistic knowledge and context. This vast pre-training dataset empowers the model to have a deeper understanding of the language, including nuanced expressions and review-related content. Consequently, when fine-tuned on review ratings, it demonstrates a greater proficiency in identifying and comprehending the subtle nuances inherent in rating assessments, thus leading to a higher accuracy rate of 64.24%.
- Furthermore, the GPT-3.5 architecture incorporates optimized transformer-based deep learning mechanisms, which are capable of capturing intricate relationships and dependencies within the training data. This results in a more sophisticated representation of review content and its correlation with rating scores. In contrast, despite the LLaMA-2 model being fine-tuned and utilizing an optimized transformer architecture, SFT, and RLHF, it may not benefit to the same extent from such a sophisticated architecture. This difference may explain its slightly lower success rate of 61.95%.
- Additionally, the GPT-3.5 model's architecture inherently facilitates better generalization, enabling it to adapt more effectively to the nuances and variations within the review data, which may arise from different sources and domains. This adaptability is a consequence of the model's extensive pre-training and fine-tuning stages.

#### **The importance of domain-specific fine-tuning**

- **Research Question 3:** How significant is the process of fine-tuning LLMs for specific tasks?

- **Research Statement 3:** The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 10.99% and 9.41% increase in predictions in comparison to their base models predictions.

Building upon the results of our research, it is essential to situate our findings within the context of how these fine-tuned models outperform their respective base models. In this context, we detect a noteworthy and substantial enhancement in prediction accuracy. The fine-tuned GPT-3.5 model showcases an impressive 9.41% increase in prediction accuracy, while the LLaMA-2 model exhibits an even more remarkable 10.99% rise in prediction accuracy when compared to their base models, respectively. In a direct comparison, it becomes apparent that the LLaMA model responds slightly more effectively to fine-tuning than the GPT model. These observations collectively emphasize the fundamental significance of utilizing fine-tuning techniques to optimize LLMs for specific tasks.

#### **The importance of data quantity and quality on fine-tuning**

- **Research Question 4:** What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?
- **Research Statement 4:** The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 1.45% and 2.12% increase in predictions when the dataset was doubled in size.

In the **Evaluation of fine-tuned LLMs for half of the dataset** section, with the aim of investigating the significance of the quantity of training data in the performance of LLMs, a new training set was randomly created, comprising 50% of the records from the base training set. Using this specific set, both models were fine-tuned, and then they were tasked with predicting star ratings for same test data in the form of product reviews.

Following a comparative analysis of the models trained with 50% of the training data and the models trained on the entirety of the training data, a substantial difference in response accuracy became evident. To be specific, the fine-tuned LLaMA-2 model exhibited a 1.45% decrease in response accuracy compared to its fully trained counterpart, while the GPT-3.5 model displayed a 2.12% reduction, respectively. These findings strongly emphasize the crucial role that the quantity of training data plays in shaping the overall performance and accuracy of fine-tuned models.

Simultaneously, research from similar studies has demonstrated that, in addition to data quantity, data quality plays a significant role in fine-tuning LLMs [108]. It is crucial to emphasize that ensuring exceptional quality in labeled data is just as important as meeting quantity thresholds. Achieving high-quality data in real-world applications requires a multifaceted approach involving data cleaning, data labeling, and rigorous data quality assessment.

#### **Cost-efficiency aspect of fine-tuning on LLMs**

- **Research Question 5:** To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?
- **Research Statement 5:** Effective prompting engineering and fine-tuning of LLMs play a crucial role in achieving improved outputs and greater cost efficiency.

It is imperative to underscore the cost-efficiency aspect of fine-tuning when deploying LLMs. To elucidate this point, we must consider the computational tokens as a unit of measurement. A descriptive output of 1,000 characters, calculated through the `tiktoken` library, corresponds to 204 tokens. These tokens are a representation of the cost incurred for executing a model in terms of the computational resources required for the model to comprehend the input prompt and generate a response for the end user.

According to the billing policy of `gpt-3.5-turbo-1106`, these 204 tokens amount to an output cost of 0.000408 USD (0.002 USD per 1,000 tokens). In contrast, a structured output, consisting of 20 characters (equivalent to 4 tokens), would incur a cost of 0.00004 USD for the operator running the GPT-3.5 model. While the difference may seem inconsequential for a single output, when applied

to a dataset with millions of records necessitating millions of outputs, the cost differential becomes substantial. The cost difference is even more significant for the fine-tuning of the LLaMA-2 model, which, in our research, required 8x A40 (Large) GPUs for 377.02 min in runtime. Particularly during the prediction phase of `llama-2-70b-chat`, due to its chat-oriented nature, responses before fine-tuning contained dialogue text, even though the prompt explicitly specified returning JSON format. This fact accentuates the critical role of fine-tuning LLMs in cost reduction.

Furthermore, cost efficiency is particularly pertinent in the context of the input prompt. Fine-tuned models have the potential to yield equivalent results to those generated by a base model, but with a more concise prompt, using the appropriate prompting engineering techniques. This holds significant implications for the economic consideration associated with deploying LLMs.

In summary, this examination highlights the economic implications of fine-tuning LLMs, emphasizing that while the cost difference may seem marginal for isolated outputs, its significance becomes pronounced when dealing with extensive datasets requiring numerous responses. Additionally, the ability of fine-tuned models to achieve similar outcomes with shorter prompts underscores the imperative nature of fine-tuning in optimizing both performance and cost-efficiency in the deployment of LLMs.

#### **Generated output structure on LLMs**

- **Research Question 6:** What impact does fine-tuning have on the generated output of LLMs?
- **Research Statement 6:** The generated output of LLMs can be enhanced through fine-tuning, with the LLaMA-2 model improving its responses by 100% following the fine-tuning process.

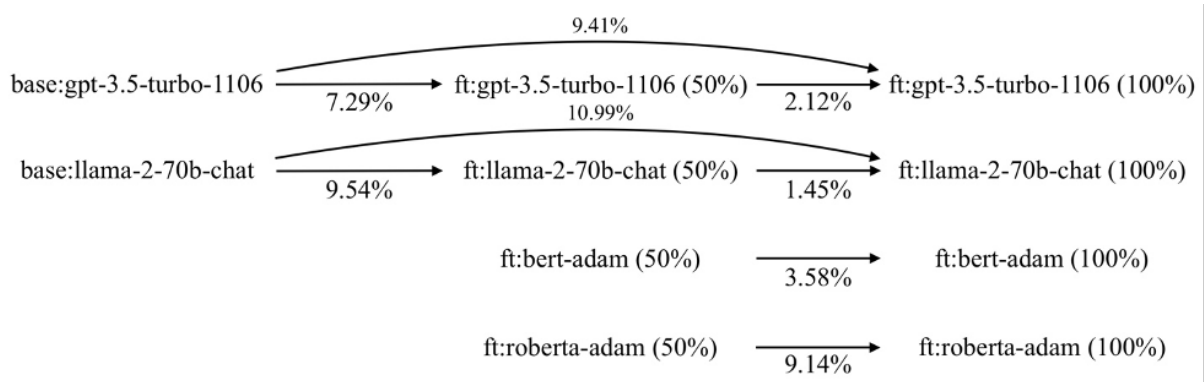
While utilizing the base model `llama-2-70b-chat`, it became apparent that, on numerous occasions, the generated output did not align with the desired JSON structure, even when explicit instructions were provided in the prompt. In contrast, the base model `gpt-3.5-turbo-1106` consistently delivered correct JSON-formatted answers from the outset. However, a pivotal turning point occurred after the fine-tuning process was applied to the LLaMA-2 model. Remarkably, the results exhibited a substantial improvement in adhering to the JSON format for all responses generated.

This noteworthy transformation in the model's behavior highlights a critical aspect of the utility of fine-tuning when working with LLMs. It serves as a compelling testament to the profound impact that fine-tuning can have on molding these models into highly effective tools for specific tasks.

The significant shift from inconsistent output to a perfect JSON format output following fine-tuning showcases the malleability and adaptability of LLMs. Fine-tuning empowers these models to not only understand the specific requirements of a task but also execute them with precision. This transformation underscores the practical importance of fine-tuning as a powerful mechanism for tailoring these models to meet specific goals, whether in structured data generation or any other specialized language processing task.

#### **Comparing fine-tuning and prediction performance: LLMs vs non-LLM NLP models**

- **Research Question 7:** Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?
- **Research Statement 7:** Overall, LLMs are more accurate in predicting product review ratings than non-LLMs after fine-tuning.



**Figure 6.** Model-wise variability in accuracy percentage of model predictions

Observing **Figure 6** depicts that the performance of the `ft:gpt-3.5-turbo-1106`, trained on 100% of the training set, is more accurate by 9.41% compared to the predictions of the `base:gpt-3.5-turbo-1106` model and 2.12% more accurate than the `ft:gpt-3.5-turbo-1106` trained on 50% of the training set.

Similarly, the performance of the `ft:llama-2-70b-chat`, trained on 100% of the training set, is more accurate by 10.99% compared to the predictions of the `base:llama-2-70b-chat` and 1.45% more accurate than the `ft:llama-2-70b-chat` trained on 50% of the training set.

In comparison to the `ft:bert-adam` model, the `ft:gpt-3.5-turbo-1106` makes more precise predictions by 5.43%, and the `ft:llama-2-70b-chat` by 3.04% (on 100% of the training set). Compared to the `ft:roberta-adam` model, the `ft:gpt-3.5-turbo-1106` makes more accurate predictions by 3.58%, and the `ft:llama-2-70b-chat` by 1.19% (on 100% of the training set).

Overall, LLMs are more efficient than non-LLMs after fine-tuning.

An interesting point observed is the percentage performance difference between models trained on 50% and 100% of the training data. The increase in accuracy for BERT and RoBERTa models is 3.58% and 9.14%, respectively, while for GPT-3.5 and LLaMA-2, it is 2.12% and 1.45%, respectively. This implies that with the increase in the training set by 1,763 samples, the predictions of BERT and RoBERTa are more accurate compared to GPT-3.5 and LLaMA-2. This finding raises additional questions about whether LLMs will continue to be equally efficient if the dataset is further increased. However, studies have shown that LLMs need more than 1,000 samples during training to understand nuances and specific patterns in a dataset [109,110]. This may suggest that if we have a small dataset, it is preferable to use BERT and RoBERTa, while LLMs become more efficient for the specific task with a much larger training set.

Using the same logic, we can also consider the cost of running and fine-tuning models for specific tasks. For training BERT and RoBERTa models on a training set of 3,520 samples, it took less than 5 min (0.41 compute units) with an A100 GPU, and the cost of 100 compute units is \$9.99 on Google Colab. In contrast, fine-tuning `llama-2-70b-chat` using the replicate API took 377 min with 8x A40 GPUs, costing \$133.47, and fine-tuning `gpt-3.5-turbo-1106` took 56 min with \$11.68. Therefore, it is up to the user to decide whether to use an LLM for higher performance, considering the cost, or choose a more economical NLP model by comparing costs and benefits. In the course of our investigation, empirical evidence substantiates the superior efficacy of LLMs in contrast to alternative NLP models when applied to the prediction of product review ratings. While the task at hand pertains specifically to regression and sentiment analysis, rendering a broad inference regarding the overarching efficiency of these particular models across classification tasks would constitute an oversimplification.

## **2.7.6. Implications and directions for future research**

### **Implications of the study**

The results of this study present both theoretical and practical implications for the fields of research related to LLMs and AI, as well as for managers, marketers, and data analysts in the e-commerce sector. Our research underscores the importance of understanding the emotions of e-commerce customers following a purchase, as well as their satisfaction levels. In the past, comprehending customer emotions was a challenging task, as models could not grasp the context behind each customer review. In this study, we recommend the use of the latest advancements in the AI field, the LLMs, as a more efficient solution for understanding the emotions behind each product review and identifying customer satisfaction. LLMs not only provide advantages to e-commerce owners, managers, marketers, and data analysts benefit, but also ensures customers receive faster responses to their feedback, thereby enhancing the e-commerce strategy's effectiveness.

Adhering to the collective findings from this study positions us to address the ultimate and most critical research question (Q8), which concerns the effective use of LLMs for evaluating product reviews. Considering the discoveries that LLMs represent a highly promising AI technology capable of providing tangible solutions across various tasks, including the assessment of sentiment in product reviews within an e-commerce context, our results, both pre and post fine-tuning, have demonstrated the considerable success of GPT-3.5 and LLaMA-2 in predicting the rating stars of reviews and identifying the words influencing user ratings.

### **Limitations and directions for future research**

Our research has highlighted the effectiveness of LLMs for tasks related to assessing reviewers' sentiments and gaining a deeper understanding of the words and context within a review that significantly impact the rating of that specific review. However, additional enhancements could be attained through both dataset and fine-tuning process improvements, as well as advancements in LLM technology. Nevertheless, within this section, we not only discuss the limitations but also emphasize certain future directions that the authors consider essential for advancing the e-commerce field and integrating LLMs into processes aimed at automating e-commerce and customer-related tasks.

#### Limitations of this study:

While the study's findings are valuable, it is crucial to consider several important aspects. First, the conclusions are derived from a randomly selected sample of 5,029 product reviews, potentially limiting their representativeness in the diverse e-commerce landscape. To ensure more comprehensive insights, future research should prioritize the inclusion of a broader and more diverse dataset. Furthermore, the study emphasizes the potential of LLMs, but highlights the need to acknowledge that their adaptability may not be universal across the e-commerce spectrum. Factors such as specific product categories, service models, and diverse customer demographics can significantly influence the effectiveness of LLMs in various contexts. In addition, the use of LLMs for scrutinizing customer sentiment raises ethical concerns, particularly related to privacy and data security. These ethical quandaries need thoughtful consideration and appropriate measures to ensure responsible use.

#### Future Directions:

Looking forward in the field of AI for e-commerce, several key pathways for advancement come into focus. Firstly, there is a crucial need for a comprehensive comparative performance assessment, meticulously evaluating how LLMs compare to other AI models to ascertain their suitability for specific e-commerce applications. Our research has highlighted the importance of evaluating the performance of LLMs, incorporating scalable increases in the training set and fine-tuning. This involves calculating prediction accuracy and the learning rate of the model for each task, allowing, through scalable fine-tuning, for precise predictions of the most efficient training set size in the future.

Additionally, the development of a robust ethical framework is paramount. This framework should offer clear and comprehensive guidelines for the responsible integration of AI into e-commerce, with a strong emphasis on the ethically handling customer data and ensuring privacy protection.

Moreover, it is imperative to delve deeply into segmentation strategies. This entails the customization of LLMs to cater to distinct customer segments and product categories, thereby enhancing their effectiveness and relevance. Furthermore, we should explore the integration of multi-modal data, encompassing text, images, and videos, into sentiment analysis to attain a more holistic understanding of customer emotions and satisfaction.

Finally, the future of e-commerce involves not only using LLMs for analytical purposes, but also harnessing their power for the personalization of recommendations and the optimization of chatbot interactions. This holistic approach aims to significantly elevate the overall user experience, leading to increased customer satisfaction and engagement.

### **2.7.7. Section Conclusion**

In conclusion, understanding consumer sentiments, purchase decisions, and ensuring unconditional customer satisfaction are vital for the sustained success of e-commerce enterprises in the face of intense competition. Technology, especially LLMs, offers robust solutions to common e-commerce challenges. In this study, the use of LLMs for understanding reviewer satisfaction post-purchase was explored. The results were highly encouraging, with LLMs, both in their base and fine-tuned forms, achieving predictive accuracy rates approaching 65% for review ratings. A notable observation was the alignment of lexical elements considered significant by LLMs in rating assessments, underscoring their ability to comprehend the importance of specific words within the context, similar to human evaluators. The use of LLMs is undoubtedly a promising tool, benefiting e-commerce owners in making strategic decisions and offering customers valuable insights that enhance their satisfaction and experience.

## ***2.8. Precision-Driven Product Recommendations Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems***

### **2.8.1. Section Summary and Significance of the Study**

The cart phase represents a pivotal juncture in e-commerce, serving as both the moment of truth and an opportune moment for upselling. Personalized product recommendations during this stage are crucial for elevating the average order value (AOV) without compromising conversion rates [111].

Evaluating recommender systems has consistently posed a challenge. Several metrics, including user surveys, accuracy metrics, user engagement metrics, precision and recall, and A/B testing, have been introduced as methods to assess recommendation systems. While several evaluation metrics are available, the subjective judgment of customers emerges as the crucial factor in determining the authentic success of a recommender system. Their discernment regarding the value of a product recommendation becomes the ultimate benchmark for the system's effectiveness. As such, striking a balance between quantitative metrics and customer satisfaction remains a paramount consideration in the ongoing refinement of recommender systems.

In this study, we propose an innovative methodology to elevate precision within product recommendation systems through the integration of advanced unsupervised ML models with the state-of-the-art GPT-4 LLM [3]. Our investigation focuses on three unsupervised models—specifically, K-Means clustering, Content-Based Filtering (CBF), and Hierarchical Clustering—all meticulously trained to enhance the accuracy of product recommendations. To evaluate the effectiveness and conduct comprehensive comparisons among these models, we harness the robust capabilities of the GPT-4 model [112].

The devised methodology incorporates the development of a user-friendly Flask-based API, strategically designed to provide e-commerce owners with a seamless solution for the integration and optimization of their product recommendation systems. Leveraging this API, e-commerce proprietors can effortlessly input their product data in CSV format, initiating an automated process that encompasses the training and evaluation of the three unsupervised models. The pivotal inclusion of the GPT-4 LLM introduces a groundbreaking dimension to our approach, enabling a sophisticated understanding and refinement of the semantic context associated with product features. This augmentation significantly amplifies the precision and relevance of the recommendations provided.

In practical application, our proposed system facilitates user interaction by allowing e-commerce owners to input a product title via the API. Subsequently, the software employs the trained models and GPT-4 LLM to deliver the most contextually fitting product recommendations. This pioneering framework not only streamlines the implementation of intricate unsupervised models but also capitalizes on the advanced natural language understanding capabilities inherent in GPT-4, resulting in refined and personalized product recommendations.

The empirical results derived from our extensive experiments underscore the superiority of employing an LLM as an evaluation tool over the time-consuming and cost-inefficient process of human evaluation. Our proposal involves utilizing the GPT-4 model as an evaluation tool to refine the precision of recommendations generated by traditional unsupervised algorithms. This is achieved through multiple rounds of training and evaluations on these models, with adjustments made to model parameters or hyperparameter tuning, all aimed at enhancing recommendation precision.

This paper significantly contributes to the ever-evolving landscape of recommender systems by introducing an efficient and scalable solution. It empowers e-commerce enterprises to optimize their product recommendation systems by seamlessly integrating state-of-the-art ML and NLP technologies.

The primary aim of this research is multifold: firstly, to assess and compare the performance of the supervised models for recommendation tasks; secondly, to introduce the GPT-4 LLM for model evaluation; and thirdly, to address specific research questions that have not been adequately answered by prior studies:

Q1: Which unsupervised recommender system algorithm demonstrates superior efficacy in product recommendation tasks?

Q2: Is the GPT-4 model capable of evaluating the predictions of traditional unsupervised models?

Q3: Can LLMs replace human evaluations?

Q4: Do NLPs have the capability to evaluate recommendations from unsupervised learning algorithms?

Q5: Can the proposed Flask-based API contribute to the accessibility and usability of sophisticated recommendation systems for e-commerce owners?

To address the aforementioned research questions, the study initiates by providing a concise literature review in **Section 2.8.2**. **Section 2.8.3** outlines the research and development methodology, encompassing the training of unsupervised models and their evaluation using the GPT model. **Section 2.8.4** delves into the results, extracting insights and formulating statements that address the research questions. Finally, **Section 2.8.5**, an alternative traditional method is presented for assessing the effectiveness of recommender systems, drawing comparisons with our proposed GPT-based evaluation tool.

### 2.8.2. Literature Review

In the ever-evolving realm of e-commerce, the curation and presentation of recommended products have traditionally rested on the shoulders of e-commerce owners, entailing a manual and intricate process across various online platforms, encompassing both product pages and shopping carts. E-commerce proprietors historically undertook the laborious task of manually selecting products believed to align with the current visitor's preferences, or proposing complementary items strategically timed just before the checkout stage [113]. This manual curation spanned a spectrum of strategies,

including cross-selling, up-selling, bundle recommendations, frequently bought together suggestions, and personalized recommendations. The associated workload for e-commerce owners in executing these strategies was undeniably colossal.

However, the landscape of e-commerce underwent a paradigm shift with the advent of ML and AI, ushering in a transformative era. This era introduced automated solutions through models and algorithms meticulously trained for the explicit purpose of handling the intricacies of recommendation systems [114]. Initially adopted by major marketplaces, these models have since evolved and adapted to augment and streamline the product recommendation processes within individual e-commerce platforms.

This literature review section embarks on an exploration of the profound evolution brought about by ML and AI in the realm of e-commerce product recommendation systems. It delves into the most renowned product recommendation strategies, elucidating how these strategies have been transformed and optimized through the integration of automated technologies. Moreover, the review aims to present the most effective algorithms tailored for each specific recommendation strategy, shedding light on the efficiency, accuracy, and adaptability of these intelligent systems in catering to the diverse needs of online shoppers.

### **Personalized Product Recommendation Strategies**

In the realm of e-commerce, strategic approaches such as cross-selling play a pivotal role in maximizing customer value and satisfaction. Specifically, complementary product recommendations form a key component of this strategy, where businesses suggest products that complement those already in the customer's shopping cart [115]. For example, if a customer adds a camera to their cart, cross-selling might involve recommending accessories like lenses or tripods. This not only enhances the overall shopping experience by providing relevant options but also opens avenues for the seller to increase the average transaction value and build a more comprehensive relationship with the consumer [116]. Thoughtful execution of a cross-selling recommender system not only boosts sales but also fosters customer loyalty by showcasing a genuine understanding of their needs [117].

Conversely, up-selling is another strategic sales tactic in e-commerce that focuses on encouraging customers to consider higher-value or superior-quality alternatives to the items in their cart [118]. By suggesting premium product alternatives, such as an advanced smartphone model with enhanced features, businesses aim to entice customers to spend more [119]. This approach not only contributes to increased revenue but also elevates customer satisfaction and loyalty by highlighting the added value associated with premium offerings [117]. Through an effective up-selling recommender system, customers are provided with an opportunity to explore superior options and enhance their overall shopping experience.

Moving on to bundle recommendations, this approach involves suggesting sets of products that include items already in the customer's cart along with additional related products [120]. For instance, if a customer selects a camera and accessories, a bundle recommendation might offer a complete photography kit with added items like a camera bag and cleaning kit. This strategy simplifies the shopping process for customers, offering a ready-made solution and often presenting a cost-saving opportunity compared to individual purchases [121]. Bundle recommender systems enhance the overall shopping experience by catering to the customer's needs in a holistic manner, contributing to increased sales and customer satisfaction [122].

Frequently Bought Together, rooted in collaborative filtering algorithms and historical purchasing patterns, suggests products commonly bought in tandem with items in the customer's cart [123]. By analyzing previous customer preferences using user profiling techniques, this strategy offers companion products or popular pairings. For example, if a customer selects a laptop, the system might recommend commonly paired accessories like a laptop bag or wireless mouse. This data-driven approach streamlines decision-making, enhances the shopping journey, and increases the likelihood of upsells and additional purchases, contributing to a more satisfying overall experience [124].

In the era of modern e-commerce, personalized recommendations stand as a cornerstone, leveraging customer data, preferences, and behavior to offer tailored and customized product suggestions [121]. This sophisticated approach analyzes diverse factors such as past purchases, browsing history, and demographic information to create a highly personalized shopping experience. By providing unique recommendations aligned with each customer's specific interests and needs, businesses aim to enhance customer engagement and satisfaction [125]. Whether suggesting similar products based on past purchases or offering personalized promotions, this strategy not only contributes to a more enjoyable and efficient shopping experience but also fosters customer loyalty through a thoughtful and individualized approach [117].

### **Product Recommendations Models and Algorithms**

In the ever-evolving realm of e-commerce, the integration of ML and AI models, particularly in recommender systems, is ubiquitous. These models play a crucial role in providing a wide array of dynamic and efficient product recommendations. Each recommendation strategy exhibits its distinctive characteristics, utilizing varying datasets, features, and objectives. Consequently, the selection of the appropriate model becomes paramount in ensuring the effectiveness of the recommendation system.

In the realm of cross-selling, the application of algorithms is instrumental in scrutinizing customer behavior and proposing complementary products [115]. The Apriori algorithm, a key player in Association Rule Mining, adeptly identifies patterns in purchased items, enabling businesses to strategically promote related products and enhance the likelihood of cross-selling success [126]. Collaborative Filtering, encompassing user-item interaction and item-item collaborative filtering, leverages collective preferences to provide personalized recommendations aligned with individual customer tastes, enriching the cross-selling potential [127].

Up-selling strategies, designed to promote higher-value products, rely on sophisticated algorithms like Decision Trees. These trees analyze customer behavior, pin-pointing factors leading to premium purchases [128]. Concurrently, ranking models such as RankNet or LambdaMART contribute by predicting and ranking products based on revenue potential, ensuring personalized recommendations that resonate with customer desires [129]. The synergy between Decision Trees and Ranking Models equips businesses to optimize up-selling initiatives and maximize revenue opportunities [128].

For crafting effective bundle recommendations, businesses employ algorithms that comprehend inherent product relationships. Market Basket Analysis and clustering algorithms prove indispensable, revealing patterns in customer transactions and organizing products into cohesive clusters [126]. These insights empower businesses to recommend entire clusters as bundles, enhancing the overall shopping experience and boosting sales through well-matched product combinations.

Frequently Bought Together recommendations capitalize on advanced algorithms, with Collaborative Filtering playing a pivotal role. By analyzing user preferences and behaviors or drawing connections between similar items, this technique suggests products commonly purchased together [127]. Association Rule Mining complements this approach, identifying relationships within transaction data and offering insights into tandem purchases [130]. The amalgamation of collaborative filtering and association rule mining enhances recommendations, providing customers with suggestions aligned seamlessly with their preferences and purchase history.

In addition, Personalized Recommendations rely on a diverse set of algorithms, prominently featuring Collaborative Filtering to recommend products based on similar user profiles (user profiling) [131]. Content-Based Filtering suggests items based on the features of previously engaged-with products [132]. Matrix Factorization techniques, including SVD and ALS, contribute by breaking down user-item interaction matrices [133]. Deep Learning Models, such as RNNs and NCF, offer complex and accurate personalized recommendations by learning intricate patterns in user behavior data [134]. The fusion of these algorithms ensures a nuanced and highly personalized shopping experience, reflecting individual preferences and enhancing user satisfaction.

Crucially, many modern recommender systems adopt hybrid approaches, combining various techniques to harness their strengths for more accurate and diverse recommendations [135,136]. The choice of algorithms hinges on the dataset, product characteristics, and the specific goals of the e-commerce platform.

### 2.8.3. Materials and Methods

The current research is centered on the implementation of cross-selling, up-selling, and personalized recommendation strategies for customers who have already added products to their basket. This encompasses the integration of recommender systems to enhance the overall shopping experience. The primary objective is to encourage customers to consider additional related products that complement those already in their cart. This strategic approach is designed to elevate customer satisfaction with the offered products and concurrently augment the overall order value and average order value (AOV), thereby fostering additional sales for the e-commerce platform. For the purpose of this research, three distinct unsupervised ML algorithms were chosen and trained on a shared dataset. Following their training, these algorithms were assigned the task of generating product recommendations for specific product titles and categories.

Each algorithm employed a unique approach, either clustering products or recommending items based on their attributes. The models aimed to provide recommendations or predictions for the most fitting product within the training set for each product in the testing set. The resultant recommendations for each product in the testing set, generated by the three trained models, were stored in a CSV file for subsequent analysis.

The selected algorithms were as follows:

- **K-means Clustering Algorithm:** It is a widely adopted unsupervised ML technique designed to organize data points into distinct groups or clusters, leveraging their shared characteristics [137]. The "K" in K-means signifies the algorithm's objective of identifying a specific number of clusters within the dataset [138]. Through iterative processes, the algorithm assigns data points to clusters and refines cluster centroids until a convergence criterion is satisfied.
- **Content-Based Filtering (CBF):** It stands as a distinctive recommendation system methodology, drawing its efficacy from the inherent attributes of items and user inclinations to curate personalized suggestions. Diverging from collaborative filtering, which relies on user-item interactions, CBF zeroes in on the intrinsic content characteristics of items and user profiles (user profiling) [127]. Items are delineated through features or metadata like keywords, genres, or pertinent descriptors. By scrutinizing users' historical preferences, the system adeptly suggests items aligning with their past choices or explicitly stated preferences [139].
- **Hierarchical Clustering:** It is recognized as a robust technique in data analysis and clustering, orchestrating the arrangement of data into a hierarchical tree or dendrogram structure based on similarities among data points [140]. This method systematically builds clusters by iteratively merging or dividing existing clusters until a comprehensive hierarchy is established. The decision-making process in hierarchical clustering, deciding whether to group or separate data points, relies on a selected distance metric like Euclidean distance or correlation [141]. The agglomerative approach starts with individual data points, progressively merging them into clusters, whereas the divisive approach begins with a single cluster and iteratively fragments it into smaller clusters [142]. This hierarchical representation provides a nuanced comprehension of relationships and structures within the dataset, unveiling insights into the diverse levels of similarity and dissimilarity between data points.

Post-prediction, the evaluation phase utilized the GPT-4 model, considering that the product suggested by the algorithms is related to the items already present in the customer's cart and could be recommended for combined purchase.

To fulfill the objectives of this study, a specific methodology was adhered to, encompassing distinct steps for training, recommendation generation, recommendations' evaluation and software

development. The ensuing subsections outline these steps to present a comprehensive and cohesive strategy for the study.

### **Dataset splitting and preprocessing**

For the research purposes, the Amazon UK Products dataset was utilized [143]. This specific dataset is available on the Kaggle platform and is distributed under the Open Data Commons Attribution License (ODC-By) v1.0. The ODC-By v1.0 license allows users to freely use, modify, and share datasets, provided proper attribution is given to the original data source. The dataset has a size of 137 MB and contains 2.2 million products. This dataset was chosen due to its origin from a reliable source, with each product in the collection offering comprehensive information such as title, category, ASIN, price, reviews, stars, and more. In this study, although all columns were retained from the original dataset, only the `title` and `categoryName` columns were used for both training and recommendations.

In the initial phase, a random selection of 10,000 products was conducted using the `train_test_split` function from the `sklearn` library [144], chosen from the pool of 2.2 million products in the dataset. Subsequently, a new CSV dataset was generated to encompass this subset of products. Care was taken to ensure that for each entry in the new dataset, the fields of product title and category were populated. Subsequently, the dataset was split into training (70%), validation (15%), and test (15%) sets, and each set was saved into separate CSV files.

Both the 10,000-product dataset and the training, validation, and test sets are available in a GitHub repository [145].

### **Training K-Means Clustering, Content-Based Filtering (CBF), and Hierarchical Clustering Models for Product Recommendations**

In this section, the methodology followed for both training and predictions/product recommendations is presented separately for each model. To expedite the results, the training and prediction phases were conducted on Google Colab using a A100 GPU [146]. The source code for each training and recommendation phase for the three algorithms is available in a `ipynb` file on GitHub [145].

#### *K-Means Clustering Training*

For the K-Means clustering training phase, the process begins by importing necessary libraries and mounting Google Drive, to access and save files. Global variables for TF-IDF transformers are defined, and a function to transform input data using pre-trained TF-IDF transformers is implemented [147].

The K-Means clustering model is trained using a combined dataset of training and validation data. The input features are extracted using TF-IDF for both the `title` and `categoryName` attributes. The number of clusters is set to 10, and the model is trained on the combined transformed data. The fine-tuned model, along with the TF-IDF transformers, is saved for future use.

The model's performance is evaluated on the validation set using the Silhouette Score, which measures the well-definedness of clusters in the data. Recommendations are made based on input data, with the CSV file being updated to include the `kMeansRecommendation` column.

To make recommendations, the model is loaded, and input data is transformed using TF-IDF transformers. The cluster is predicted for the input data, and products within the same cluster are identified. The cosine similarity between the input product and each product in the cluster is calculated, and recommendations are sorted based on these similarity scores. The top N recommendations are then returned.

Finally, a function iterates through all the products in the test set. It retrieves recommendations, selects the top N recommendations for each product, and stores the first one in the `kMeansRecommendation` column.

### *Content-Based Filtering (CBF) Training*

In the outlined procedure for training and generating product recommendations using Content-Based Filtering (CBF), the initial step involves mounting Google Drive to access and save files. The Python script utilizes the `pandas` library to handle data frames, the `joblib` library for file I/O operations, and `scikit-learn` for NLP tasks, employing the `TfidfVectorizer` to transform product titles and categories into feature vectors [148]. A separate function is created to facilitate the transformation of input data based on pre-trained TF-IDF transformers, and the main training function is then employed to train these transformers using the provided training data.

The core recommendation process is encapsulated in a function that calculates cosine similarity between the input product and those in the training data, sorts them by similarity scores, and returns a list of recommended product titles. Finally, another function utilizes the trained transformers to update a CSV file with recommendations, incorporating the `cbfRecommendation` column for each product. This function iterates through the existing data, applies the recommendation algorithm, and updates the CSV accordingly. The entire process is orchestrated to enhance product recommendations based on the Content-Based Filtering approach.

### *Hierarchical Clustering Training*

For Hierarchical Clustering, the initial steps involve importing necessary libraries and mounting Google Drive to access and save files. Two global variables, `title_tfidf` and `category_tfidf`, are declared to store pre-trained TF-IDF transformers. The `transform_input` function is then defined to transform input data using these pre-trained transformers. The training process begins by combining the training and validation datasets and extracting features using TF-IDF for both the product title and category. The chosen number of clusters is set to 2, and an Agglomerative Clustering model is trained on the combined transformed data. The model, along with the TF-IDF transformers, is saved for future use. Evaluation is performed using the Silhouette Score.

To generate product recommendations, the `recommend_product` function is designed to predict the cluster for the input data, identify products in the same cluster, calculate cosine similarity scores, and sort products based on these scores. The top N recommended titles are then returned, and the first of them is selected. The `get_recommendations` function loads the pre-trained model, makes recommendations for a specific row in the dataset, and updates a CSV file with the recommendation information. The entire process is executed on a training and validation dataset, and the recommendations are stored in the `hierarchicalRecommendation` column in the test set.

### **Zero-Shot Evaluation Methodology using GPT-4 model for Assessing Product Recommendations**

The primary objective of our research was to discover an innovative approach for evaluating the efficiency of the specific algorithms in product recommendation beyond the conventional methods presented in earlier studies. Many research endeavors comparing models in similar tasks often employ the similarity score and other evaluation metrics, which utilize the `SequenceMatcher` from the `difflib` module to calculate the similarity ratio between the product title and the recommended product title [149,150]. In this study, we propose the use of the GPT-4 model for evaluating the effectiveness of models in product recommendations. As mentioned in **Section 2.3**, the newly introduced GPT-4 model by OpenAI is an LLM that exhibits numerous applications in various domains and tasks.

At this stage of the research, we possess a comprehensive CSV file containing product recommendations for each item in the test set. At this juncture, we construct a prompt designed to prompt the GPT model to evaluate the recommendations made by the three models for each product in the test set. For every product and each recommendation, the GPT model was tasked with assessing the recommendation with a binary rating: 1 indicating a high likelihood of customer purchase, and 0 signifying a low probability of purchase. In addition to the binary evaluation, to gain further insights from the GPT model's assessment, the model was prompted to provide textual justification for its

decision. Both the evaluation scores and the accompanying textual justifications were stored in the CSV file. After multiple attempts, the selected prompt capable of simultaneously performing both tasks is presented in **Figure 7**.

```
conversation = []
conversation.append({'role': 'system',
                    'content': "You're assisting a customer as a salesperson, and they've added" +
                               " the product with the title [" + title + "] to their basket."})
conversation.append({'role': 'user',
                    'content': "Evaluate whether it's a good idea to recommend adding the product" +
                               " with the title [" + recommendation + "] as an extra." +
                               " Provide the reason in JSON format why the customer might accept" +
                               " your offer {\"accept\": 1, \"reason\": 'add here the reason for accept'}" +
                               " or the reason they could decline your offer" +
                               " {\"accept\": 0, \"reason\": 'add here the reason for decline'}."})
```

**Figure 7.** GPT prompt for evaluating product recommendations

The dialogue in **Figure 7** is structured as a series of dictionaries, with each dictionary corresponding to a turn in the conversation. The initial dictionary contains the system's prompt, functioning as a salesperson proposing a product recommendation aligned with an item already present in the customer's cart. The second dictionary contains the user's prompt, in which we ask the system to assess the product suggested by the unsupervised algorithms.

For the execution of the GPT-4 model, the official OpenAI API was utilized [151]. To achieve the desired JSON format as the return from the API, the prompt was carefully designed to provide clear instructions to the GPT model regarding the expected structure of the JSON format. In the majority of cases, the results were in the correct structure. However, for instances where the GPT model returned additional text, an additional function was implemented to locate the JSON within the text, convert it into a suitable dictionary, and appropriately process and store it in the CSV file.

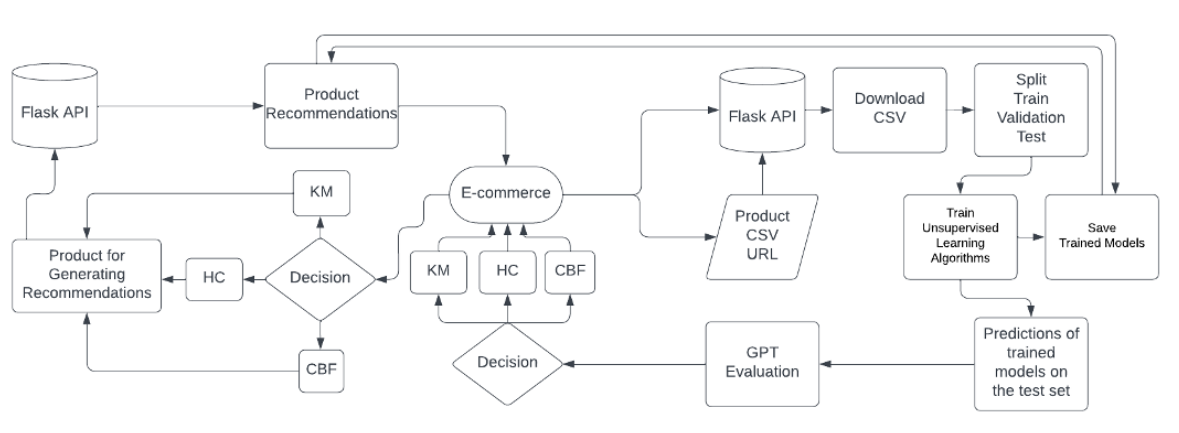
### **Development of a Flask-Based API Software for Integrated Training and Deployment of Product Recommendation Models**

Our research aims to empower e-commerce owners by facilitating the integration of advanced ML algorithms into their recommendation systems, even without prior experience in ML or AI. While there are existing automated tools or extensions for e-commerce platforms that offer recommendation system capabilities, some may rely on simplistic ML algorithms that lack customization for specific product sets. Moreover, these algorithms may not be easily trainable on low-budget web servers.

In contrast, our proposed solution, a flask-based software, enables any webmaster with basic programming knowledge to deploy it on a web server (such as PythonAnywhere hosting). By leveraging the API features, the webmaster can seamlessly connect their e-commerce platform with our recommendation software. The distinct advantage lies in the ability to train the algorithm on the specific products of the e-commerce site, ensuring tailored product recommendations.

During the development phase, we consolidated functions previously used on Google Colab, organized them into classes, and encapsulated them within a flask application. Once set up on a web server, the webmaster can initiate the process by providing the e-commerce products' URL in a CSV format through the API. The software downloads the CSV, conducts training and GPT evaluation, and identifies the algorithm that delivers the most effective product recommendations. Subsequently, the webmaster can utilize the same API to make calls, supplying a product title, and the software will predict the best matching products likely to be purchased by the customer.

To further explain API deployment and interaction with the user, a flow chart was created using Lucidchart software [**Figure 8**]. This block diagram describes both the training phase and product recommendation phase.



**Figure 8.** Training and Product Recommendations using the Flask API

Our software can be further enhanced to provide additional functionalities for the webmaster. Moreover, expanding the training of algorithms for the recommendation task offers the webmaster a broader range of options, whether for prediction speed or improved recommendation outcomes. The source code of the software is available on GitHub in an open-source format under the MIT license [145].

### Critical Factors Shaping Software Effectiveness in Collaborative and Real-Time Environments

In the realm of collaborative projects or real-time scenarios requiring human interaction, the effectiveness of the software discussed in this study hinges on several key factors:

- **Seamless Integration:** The Flask-based API offers an intuitive interface for integration, streamlining the process for developers and e-commerce proprietors to seamlessly incorporate and deploy the model in real-time applications.
- **Automated Training and Evaluation:** The software automates the training and evaluation processes, sparing users the intricacies of managing ML algorithms, model validation, and assessment.
- **Enhanced Natural Language Understanding with GPT-4:** Leveraging GPT-4 for model evaluation harnesses its advanced natural language understanding capabilities, fostering effective human interaction. GPT-4's adeptness at refining the semantic context of product features has the potential to elevate the precision and relevance of product recommendations.
- **Scalability and Efficiency:** The software stands out as an efficient and scalable solution, tailored to handle varying levels of structured data specific to e-commerce needs. This adaptability is crucial for collaborative projects and real-time scenarios.
- **User-Centric Experience:** In real-time scenarios, the success often hinges on user experience. The model generates personalized and effective recommendations, enhancing the likelihood of positive user reception.

While the proposed software model exhibits promising features for collaborative projects and real-time scenarios, particularly in the context of e-commerce, practical implementation and user feedback are indispensable for a comprehensive assessment of its performance in such environments.

#### 2.8.4. Research Results

In the preceding sections, we introduced three distinct unsupervised learning algorithms tailored for product recommendations, with a specific focus on in-cart product recommendations. Through training on both a given training set and a validation set, these algorithms were meticulously trained to provide optimal predictions and recommendations for the test set.

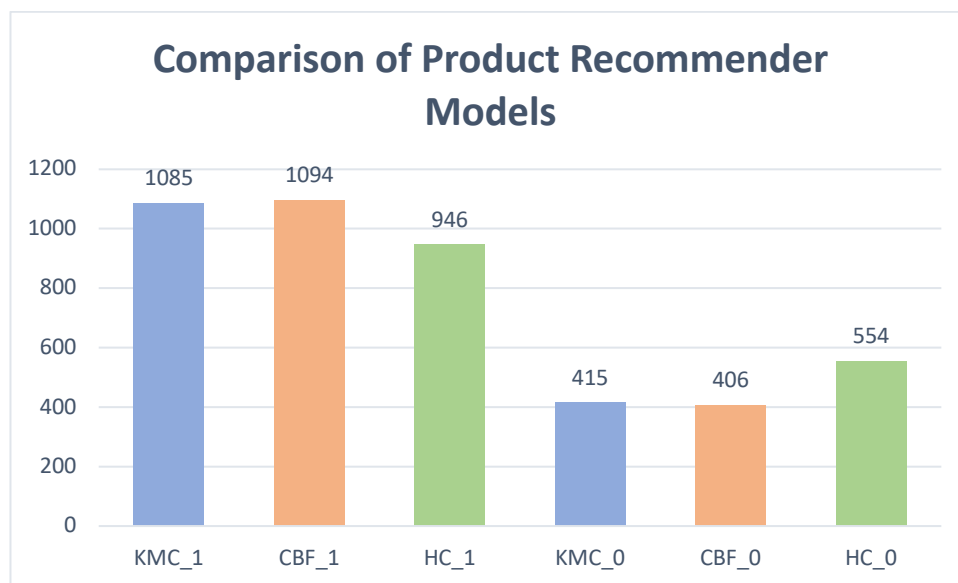
Building upon this, we propose an innovative approach in which the GPT-4 model takes on the role of evaluating the effectiveness of recommendations generated by the trained models. This

evaluation involves assigning a binary rating along with a contextual justification for the given rating. The subsequent section unveils the outcomes of this evaluation, addressing the research queries outlined in the introduction section.

**Comparing unsupervised learning algorithms trained for product recommendation in e-commerce**

- **Research Question 1:** Which unsupervised recommender system algorithm demonstrates superior efficacy in product recommendation tasks?
- **Research Statement 1:** The Content-Based Filtering (CBF) and k-means clustering trained models demonstrate higher accuracy in in-cart product recommendation tasks.

In accordance with the **Training K-Means Clustering, Content-Based Filtering (CBF), and Hierarchical Clustering Models for Product Recommendations** section, three unsupervised models underwent training on a shared dataset consisting of 7,000 products, validated with a set of 1,500 items. Following the training phase, the models were tasked with generating product recommendations from the training set for each of the 1,500 products in the test set. The resulting predictions were recorded in a CSV file, and the evaluation process was handed over to the GPT-4 model. The GPT-4 model evaluated the recommendations, assigning a binary rating on a scale from 0 to 1. A rating of 1 indicated a high likelihood of customer purchase based on the recommendation, while a rating of 0 signified a low probability of purchase. It is crucial to note that the focus of this evaluation was on in-cart product recommendations, where the customer already has at least one product in their basket, and we propose supplementary products likely to be added to their cart. The outcomes of the GPT-4 model evaluations are depicted in **Figure 9** and **Table 3**.



**Figure 9.** GPT-4 Evaluations of Unsupervised Learning Trained Models

**Table 3.** Efficiency of unsupervised learning-trained models in recommender systems

	0	1	Total
KMC Evaluation	415	1085	1500
CBF Evaluation	406	1094	1500
HC Evaluation	554	946	1500

**Figure 9** reveals that the Content-Based Filtering (CBF) trained model outperforms the others, achieving a commendable 72.93% accuracy, equivalent to 1,094 correct recommendations out of 1,500

products assessed by the GPT model. Following closely, the k-means clustering trained model achieved a 72.33% accuracy, providing 1,085 correct recommendations for the products in the test set. In contrast, the Hierarchical Clustering trained model exhibited a lower success rate, accurately recommending 946 out of the 1,500 test set products, corresponding to a 63.07% accuracy.

The GPT model evaluation underscores the superior performance of the Content-Based Filtering (CBF) trained model, likely attributed to its more sophisticated nature. The k-means clustering trained model closely follows, with a negligible distance from CBF, indicating high accuracy in its recommendations.

It's noteworthy to emphasize that achieving an accuracy of 70% is considered significant in recommendation tasks.

#### Evaluation of models using the GPT-4 model

- **Research Question 2:** Is the GPT-4 model capable of evaluating the predictions of traditional unsupervised models?
- **Research Statement 2:** The GPT-4 model exhibits significant potential in evaluating predictions from unsupervised models.

Upon delving into our research findings, it is imperative to highlight the effectiveness of the GPT-4 model as an evaluation tool. This model not only assesses trained models in a binary format but also provides insightful justifications for its ratings. **Table 4** presents some of the ratings assigned by the GPT-4 model along with their corresponding justifications.

**Table 4.** In-cart product recommendations and GPT model evaluation

In-cart Product	Recommendation Product	Model	Binary Rating	Justification
Holibanna Human Skull Teaching Model Lifesize Human skeleton Anatomical Head Bone Model Halloween Decoration	267Pcs Molecular Model Organic Inorganic Structure Kit, Atoms Bonds Orbitals Chemistry Structure Kit Atom Link Model, Educational Molecule Set for Teachers Students Student Chemistry Learning	HC	0	The customer is purchasing a human skull teaching model for Halloween decoration which suggests they are interested in anatomical models rather than chemistry models. Recommending a molecular model kit might not align with their current purchase intention.
Hellyhansen Women's Crew Sweatshirt Women's Sweatshirt	VOIO Men'S Hoodies, Warm Sweatshirt Crew Neck Print Sweatshirt Christmas Tree Plaid Print Sweatshirt 3D Pullover Suitable Autumn Winter Sweatshirt	CBF	0	The recommended product is for men, while the customer has added a women's sweatshirt to their basket. It's unlikely that the customer would be interested in purchasing a product that is not suitable for their gender.
2 Pack Silicone Foldable Cups BPA-Free Collapsible Travel Mugs 270ML Portable Reusable Cup Set with Plastic Sealing Lid for Outdoor Picnic Camping Hiking Traveling (Grey & Blue)	Light my Fire Camping Mess Kit - Picnic Set 6 Pcs - Lunchkit - Camp Mess Kit - Reusable Camping Plates and Bowls Set - BPA Free - Microwave & Dishwasher - Camping Picnic Set	KMC	1	The customer might accept adding the Light my Fire Camping Mess Kit because it complements the Silicone Foldable Cups. The Mess Kit includes reusable camping plates and bowls, which would be useful for outdoor picnics, camping, and hiking. Additionally, the fact that it is microwave and dishwasher safe makes it convenient for the customer to use and clean. Overall, the Camping Mess Kit enhances the customer's outdoor dining experience and provides them with a complete set of cookware.

<sup>1</sup> KMC: K-Means Clustering, HC: Hierarchical Clustering, CBF: Content-Based Filtering

Upon examining **Table 4**, it becomes evident that the GPT-4 model possesses significant potential as an evaluation tool. Its unique capabilities extend beyond accurately assessing predictions from other models; as an NLP model, it can also articulate justifications, effectively putting itself in the customer's shoes. By adopting our novel approach, data analysts can enhance their unsupervised recommender system algorithms, enabling more precise and insightful predictions.

#### **Comparison of LLMs evaluation and Human evaluation**

- **Research Question 3:** Can LLMs replace human evaluations?
- **Research Statement 3:** LLMs present a more cost-effective and time-efficient alternative to human evaluations, but they cannot surpass human prowess.

This research inquiry leans more towards theory than experimentation. NLP Models, including LLMs, exhibit the capability to comprehend the meaning behind human-provided text through context, and they can make evaluations akin to humans. However, human evaluations stem not only from context but also from life experiences, a dimension challenging for LLMs to replicate. Therefore, while LLMs can be utilized for evaluations in product recommendation tasks, they do not surpass humans in evaluative prowess.

Human evaluations, though effective, come with a high cost and presume that the human, often the webmaster in this context, possesses unlimited time for the task. In contrast, LLM evaluations can offer a more cost-effective and time-efficient alternative. In conclusion, LLMs may not match the success rate of human evaluations, but they can execute similar tasks swiftly, saving both time and money for the webmaster.

#### **Natural Language Models Evaluations on Unsupervised Learning Algorithms**

- **Research Question 4:** Do NLPs have the capability to evaluate recommendations from unsupervised learning algorithms?
- **Research Statement 4:** NLPs, specifically LLMs like the GPT-4 model, have proven through our study to be effective tools for evaluating product recommendations generated by unsupervised learning models.

Based on our research findings, it is evident that NLP models, particularly the GPT-4 model, demonstrate remarkable capabilities in performing diverse tasks, including the evaluation of product recommendations generated by unsupervised models. These evaluations exhibit high accuracy in both binary ratings and justifications.

A key question that arose during our research is why we persist in using unsupervised recommender system algorithms like clustering and CBF for product recommendation tasks instead of leveraging an LLM for potentially more efficient results. While an LLM might indeed excel in a comparable task, it's crucial to consider that incorporating an LLM introduces costs, whether in terms of hardware or API expenses. In contrast, evaluating the unsupervised recommender system algorithms presented in this study incurs minimal hardware costs for product recommendations.

Furthermore, our study proposes utilizing the evaluation capabilities of GPT-4 not only for providing ratings to discern which product recommendation algorithm is more efficient. Data scientists can leverage this by multi-training and evaluating these models, adjusting parameters or using hyperparameter tuning to enhance their precision, potentially exceeding the 70% accuracy reported in this study.

#### **The contribution of the software to e-commerce**

- **Research Question 5:** Can the proposed Flask-based API contribute to the accessibility and usability of sophisticated recommendation systems for e-commerce owners?

- **Research Statement 5:** The software proposed in this study has the potential to enhance both the product recommendation capabilities of e-commerce and the resulting benefits from additional sales, thereby providing high accessibility for e-commerce owners aiming to improve their recommendation system and, consequently, their e-commerce sales.

The resolution to our final research question lies in the hands of e-commerce owners who opt to employ our software to elevate their e-commerce platforms' product recommendation capabilities. The effectiveness of our software can be gauged by the additional sales potential it unlocks for e-commerce businesses. This sophisticated software not only saves time and is cost-efficient but also proves highly effective, combining the insights of an LLM with the sophisticated nature of unsupervised recommender system algorithms. Undoubtedly, our proposal stands to benefit e-commerce owners, whether by saving time on manually adding product recommendations or by enhancing the effectiveness of in-cart product recommendations.

### 2.8.5. Recommender Systems Evaluation and Discussion

In **Section 2.8.4**, the research outcomes are presented, addressing the research questions and extracting valuable research statements. This section is dedicated to assessing the recommender systems developed in this research through a traditional approach. The outcomes of this evaluation will be juxtaposed with those of the GPT-Powered evaluation tool proposed in this study, shedding light on how our assessment tool attains its binary ratings and how closely its evaluations align with the traditional approach. Following this, a discussion ensues regarding the limitations and inherent risks associated with our proposed tool.

#### Assessing Recommender Systems Effectiveness

Assessing the effectiveness of a recommender system involves gauging how well the recommended products align with user preferences or actual behavior. Various evaluation metrics, such as user surveys, accuracy metrics, user engagement metrics, precision and recall, A/B testing, and more, are commonly used. However, most of these methods require human evaluation. For instance, user surveys necessitate asking users about their satisfaction with recommended products, accuracy metrics require user ratings for recommended products, user engagement metrics involve tracking click-through rates and time spent on recommended product pages, precision and recall require feedback in binary form, and A/B testing involves randomly selected human evaluators assessing product recommendations.

Our proposed methodology introduces an alternative using GPT-4 as an evaluation tool. GPT-4 acts as a virtual human evaluator, assigning binary ratings based on its assessment of whether the product recommendation from the unsupervised model is more likely (1) or not (0) to be purchased by the customer. While the ratings are binary, the precision and recall technique cannot be applied due to its requirement of two columns—one with true values and one with predicted values.

The most effective means of evaluating both the recommendations of the unsupervised models and the performance of the GPT-4-powered evaluation tool is undoubtedly through human evaluation. However, with 1,500 rows of test data and 7,000 rows of training data, conducting human evaluation becomes time-consuming. An alternative is to assess the precision without human interference by calculating the similarity score between the product title and the recommended product title. This score possesses the ability to evaluate recommendations produced by unsupervised trained models while concurrently assessing the effectiveness of evaluations conducted by the GPT-4 model. Moreover, the similarity score can be considered a practical alternative to the GPT-4-powered evaluation tool suggested in this research.

For our study, we employed the cosine similarity measure to evaluate the effectiveness of our recommender systems. This methodology uses the `TfidfVectorizer` to convert the text into TF-IDF vectors and then calculates the cosine similarity between these vectors.

The similarity scores were calculated between the title and the recommendations made by the unsupervised trained models, and the results were stored in different columns for each of our recommendation systems (K-Means Clustering, Hierarchical Clustering, and Content-Based Filtering). Descriptive statistics for each new column have been computed and are showcased in **Table 5**.

**Table 5.** Descriptive statistics for the similarity scores in each recommendation system

Metrics	KMC	HC	CBF
count	1500	1500	1500
mean	0.315991	0.228553	0.304138
std	0.203004	0.163804	0.200361
min	0.000000	0.000000	0.000000
25%	0.151775	0.103950	0.145600
50%	0.276650	0.190900	0.259550
75%	0.440075	0.310575	0.426925
max	1.000000	0.958000	1.000000

<sup>1</sup> KMC: K-Means Clustering, HC: Hierarchical Clustering, CBF: Content-Based Filtering

Interpreting the findings from the descriptive statistics presented in **Table 5**, all three recommendation systems consistently display a count of 1500, indicating that similarity scores were calculated for the entire dataset. The mean similarity scores for the K-Means Clustering (KMC), Content-Based Filtering (CBF), and Hierarchical Clustering (HC) recommendation systems are 0.316, 0.304, and 0.229, respectively. Higher mean values suggest that, on average, recommendations moderately resemble the original titles.

Examining standard deviations, KMC and CBF exhibit similar values (around 0.203 and 0.200, respectively), implying moderate variability in scores, while HC shows a slightly lower standard deviation (around 0.164), suggesting relatively less variability. Minimum similarity scores are 0 for all systems, indicating instances with no similarity between titles and recommendations, while maximum scores of 1 suggest perfect alignment in certain cases.

Based on the descriptive analysis, it is evident that KMC and CBF perform closely and significantly outperform HC. As per **Table 3** and **Figure 9**, our GPT-Powered evaluation tool indicates that CBF and KMC are closely aligned in precision, with approximately 72.93% for CBF and 72.33% for KMC. In contrast, the HC recommendation system exhibits a lower precision of about 63.07%. These results suggest that our proposed methodology closely aligns with traditional approaches, such as similarity scores.

In our evaluation methodology, CBF is slightly more precise in recommendations than KMC by about 0.6%. Using similarity scores, KMC appears slightly more precise than CBF by 1.2%. Both evaluation tools underscore that the HC recommendation system is significantly less precise, with CBF and KMC consistently recommending the best products more likely to be purchased by the customer.

As mentioned earlier, the evaluation methodology proposed in this study closely aligns with traditional approaches like similarity scores to determine which model has the greatest potential for generating product recommendations likely to be purchased by a customer.

Upon reviewing **Table 5**, it becomes evident that the similarity scores appear to be quite low. It is important to note that a score of 1 indicates perfect alignment, while 0 signifies no alignment. In contrast, our proposed methodology demonstrates success percentages ranging from 63.07% to 72.93%. This observation highlights significant disparities in how the evaluation tools assess recommendations.

Unlike similarity scores, which attempt to measure the likeness between a given product and its recommended counterpart in vector terms, our GPT-4 evaluation tool takes a different approach. Instead of evaluating textual similarity, it comprehends the meaning behind each product title and recommendation within the context, leading to a more thoughtful evaluation.

From a customer perspective, a perfectly aligned recommendation product already in their cart may not be the optimal suggestion. Rather, understanding the nuanced meaning of the product title

and proposing a complementary product, even if not perfectly aligned, proves to be a superior recommendation. This approach holds greater potential for enticing customers to make a purchase by offering supplementary products that enhance their overall shopping experience.

The primary objective of this study is to introduce an alternative evaluation tool for recommender systems, highlighting the effectiveness of LLMs and particularly the GPT-4 model. Undoubtedly, to comprehensively assess whether our proposed methodology surpasses traditional approaches, human evaluation is deemed necessary. It is crucial to note, however, that the understanding capabilities of LLMs hold potential across various domains, making them a versatile evaluation tool applicable in diverse contexts and industries.

### **Threats To Validity**

The envisioned incorporation of GPT-4 as an evaluation tool for assessing unsupervised learning algorithms in recommender systems holds promise, but its implementation comes with inherent limitations and risks.

- **Data Bias and Generalization Issues:** The efficacy of the model hinges on the quality and diversity of the training data. Biased or insufficiently varied data provided by e-commerce owners may result in skewed recommendations that lack generalizability to a broader audience.
- **Overfitting:** The integration of multiple models, particularly with a substantial language model like GPT-4, raises the risk of overfitting. This occurs when the model excels on training data but struggles to generalize to new, unseen data.
- **Computational Resources:** The use of the GPT-4 model in the proposed recommender system could incur significant API costs for both fine-tuning and predictions, depending on the dataset size. This presents challenges for smaller e-commerce platforms with limited budgets.
- **Interpretability:** Complex integrated models may lack interpretability, making it difficult to comprehend how the model generates specific recommendations. This lack of transparency can be a concern for businesses that prioritize understanding the decision-making process.
- **Privacy Concerns:** The advanced natural language understanding capabilities bring forth privacy concerns, especially if sensitive or personal information is inadvertently captured and utilized in the recommendation process. The software developed relies on trust in the data provided by e-commerce owners through the API, without evaluating potential sensitivity.
- **Maintenance and Updates:** Keeping the recommendation system current with the latest data and ensuring compatibility with evolving e-commerce platforms poses a challenge. Regular maintenance and updates are imperative, considering the possibility of GPT-4 being deprecated in the future or changes in API functions and Python libraries.
- **User Engagement and Satisfaction:** While experimental results may suggest the effectiveness of the proposed methodology, it is essential to acknowledge that the GPT-powered evaluation tool comprehends the meaning behind each word, diverging from reliance on simple word similarity for its evaluation outcomes. Consequently, it becomes imperative to factor in user satisfaction and engagement. Users may not consistently prioritize the utmost accuracy in recommendations, underscoring the importance of striking a balance between precision and accommodating user preferences.

In conclusion, while the proposed integration of GPT-4 as an evaluation tool signifies a groundbreaking approach, addressing and mitigating these potential limitations and risks is vital for a successful real-world implementation.

### **2.8.6. Section Conclusion**

In summary, this study presents an innovative methodology that integrates unsupervised learning models, such as K-Means clustering, Content-Based Filtering (CBF), and Hierarchical Clustering, with the cutting-edge GPT-4 LLM to elevate e-commerce product recommendation systems. The training of these models, utilizing both the training and validation sets, involves

generating recommendations for a designated test set. To assess the models' recommendations, the GPT model assigns ratings ranging from 0 to 1, where 1 signifies a high likelihood of customer purchase based on the recommendation, and 0 indicates a low probability of purchase. Notably, the GPT model also provides justifications for its ratings within the same prompt.

The study's outcomes reveal compelling results, with the CBF achieving an accuracy of 72.93%, the K-Means clustering trained model achieving 72.33%, and the Hierarchical Clustering trained model achieving 63.07% in product recommendation evaluations conducted by the GPT model. Additionally, it is observed that LLMs, particularly the GPT-4 model, effectively evaluate recommender system algorithms' predictions. Consequently, the central proposition of this study advocates for the utilization of LLMs as an evaluation tool, especially in the context of product recommendation tasks. While acknowledging that GPT models cannot replace human intelligence in evaluation tasks, their demonstrated effectiveness underscores their valuable contribution, particularly in the e-commerce domain.

To translate this knowledge into practical application, we have developed a user-friendly Flask-based software with an easily installable API. This tool is designed to support e-commerce owners in enhancing their product recommendation systems, consolidating the synergy of unsupervised recommender system algorithms and advanced language understanding offered by GPT-4 LLM, thereby providing an actionable solution for businesses seeking to optimize their customer engagement strategies.

## ***2.9. GPT-Powered Software for Identifying Sustainable Product Features***

### **2.9.1. Section Summary and Significance of the Study**

In today's global context, sustainability has emerged as a crucial concept with significant implications for our planet and future generations. Sustainability encompasses a holistic approach to addressing the pressing environmental, social, and economic challenges we face [152]. It emphasizes the need to meet the needs of the present without compromising the ability of future generations to meet their own needs [153]. The significance of sustainability lies in its potential to mitigate climate change, protect natural resources, promote social equity, and foster long-term economic prosperity [154]. By recognizing the interconnectedness of these pillars, sustainability encourages responsible consumption and production patterns, innovative solutions, and collective action to create a more resilient and harmonious world [155]. As we navigate complex issues such as climate change, biodiversity loss, and social inequality, adopting sustainable practices and policies becomes imperative for a sustainable and thriving future for humanity and the planet we call home.

Each year, an increasing number of initiatives related to sustainability, both promotional and operational, come to the forefront, with numerous applications supporting the overall effort [156]. Technology, particularly AI, can aid in the promotion and development of sustainability by providing appropriate tools to businesses and individuals, enabling them to engage in and facilitate more sustainable purchases and sales [157]. However, both consumers and developers of these specific AI tools must exercise great caution regarding the limitations, potential unintended consequences, and ethical implications associated with utilizing AI for sustainability purposes.

This study presents an innovative software solution aimed at addressing sustainability challenges through the utilization of a GPT (Generative Pre-trained Transformer)-powered application. The primary objective of this software is to automate the identification of sustainable product attributes from product descriptions, titles, and product specifications. By harnessing the capabilities of NLP and ML, the software effectively categorizes products based on their sustainability characteristics. Through the analysis of textual information, it extracts crucial indicators of sustainability, such as eco-friendly materials, energy efficiency, recyclability, and ethical sourcing. By automating this process, the software eliminates the need for manual assessment, streamlining the evaluation of product sustainability. This software implementation is facilitated through an easily installable Chrome app,

which informs users about the sustainability features based on the product description and title, without requiring user intervention. Significantly, this software empowers consumers to make informed and sustainable purchasing decisions, while also assisting businesses in showcasing their sustainable offerings.

The present research extends beyond the mere development of software; rather, it delves into the evaluation of the suitability of the GPT model in identifying and assessing the sustainability of products using a real-world dataset. Four pivotal hypotheses are introduced, which will be substantiated through the outcomes of the experiments:

- **Hypothesis 1. Research Objective:** The study seeks to evaluate the accuracy of GPT model responses.
- **Alternative Hypothesis:** The GPT model produces accurate responses when appropriate prompts are employed.
  
- **Hypothesis 2. Research Objective:** The study aims to examine the format of responses generated by the GPT model.
- **Alternative Hypothesis:** The format of GPT model responses is accurate and aligns with the explicitly defined structure in the prompt.
  
- **Hypothesis 3. Research Objective:** The study aims to assess the effectiveness of the GPT model in identifying sustainable features based on product titles and descriptions.
- **Alternative Hypothesis:** The GPT model is capable of evaluating the sustainable features of a product when provided with its title and description.
  
- **Hypothesis 4. Research Objective:** The study aims to evaluate the effectiveness of the GPT model in identifying product sustainable features, particularly when product titles and descriptions are less descriptive.
- **Alternative Hypothesis:** The GPT model can successfully retrieve sustainable features of a product even when the provided title and description lack sufficient descriptive information.

Through these hypotheses, we intend to gain insights into the performance and potential applications of the GPT model in the context of sustainable product identification. By exploring its accuracy, format, and effectiveness under various conditions, we aim to contribute valuable findings to the field of NLP and its practical implications in sustainability assessment.

The research is organized as follows: In **Section 2.9.2**, an in-depth analysis of the three pillars of sustainability and their interplay with products and their corresponding attributes is presented, drawing upon a comprehensive review of current academic literature. Moreover, it rigorously investigates the predominant sustainable product attributes that align with these pillars, offering valuable insights into the prevailing trends and considerations within this domain. In **Section 2.9.3**, the study provides a comprehensive presentation of the research objectives, significance, and contribution to the domain of sustainability research. Additionally, it expounds upon the rigorous methodology encompassing data collection and preprocessing, as well as the process of model evaluation. Proceeding to **Section 2.9.4**, the research delves into diverse aspects, including the detailed depiction of the experimental setup and results, an analysis of user interaction, insights into software implementation and system design, and illustrative case studies and examples. Furthermore, the section also presents the results derived from the research, which effectively validate the proposed research hypotheses. The final section, **Section 2.9.5**, offers a comprehensive analysis of the authors' reflections concerning potential future enhancements of the software. Furthermore, it conducts an exploration of the ethical considerations associated with the application of the GPT model in the context of sustainability assessment tasks.

## 2.9.2. Sustainability Literature Review

The realm of e-commerce has been continuously evolving, exerting a direct impact on consumer purchasing habits [158,159]. While consumer choices have expanded, the corresponding legal frameworks governing consumer protection have not progressed at an equal pace [160]. Aspects such as consumer protection of privacy, misinformation, product quality, environmental impact, and related concepts necessitate addressing to safeguard consumers in the realm of e-commerce. Apart from legislative measures, educating consumers about responsible e-commerce practices is of paramount importance. AI, in turn, holds the potential to empower consumers to make informed decisions, whether pertaining to product quality or considerations of sustainability and environmental impact.

Sustainability is not an individual effort; it requires collective action and collaboration across all sectors of society to transform our consumption patterns and create a sustainable future [152,154]. Both production companies and retail businesses, as well as consumers, need to collectively change their mindset regarding sustainability [156]. Before designing and manufacturing a product, production companies should always consider how it will bring benefits or at least have a neutral impact on the environment and the economic and social aspects. Simultaneously, retail businesses should have sustainability awareness in order to inform and promote sustainable products to their customers, educating them on how an individual sustainable purchasing decision can make a difference [160,161]. Furthermore, consumers should develop a sense of sustainability and contribute to a more sustainable future by actively supporting and prioritizing sustainable products [162]. Through their support of sustainable products, individuals can effectively convey a clear message to industries and policymakers that they prioritize long-term ecological balance over short-term profits.

Sustainability can be defined as the practice of meeting the present generation's needs without compromising future generations' ability to meet their own needs [153,160]. It is a comprehensive approach that recognizes the interconnectedness of environmental, social, and economic aspects in achieving long-term well-being [163]. The environmental pillar of sustainability focuses on preserving and protecting natural resources, mitigating climate change, and promoting biodiversity conservation [155,164]. The social pillar emphasizes the importance of social equity, human rights, and community well-being, aiming to ensure that all share the benefits of development [155,165]. The economic pillar recognizes the need for sustainable economic growth that promotes prosperity, job creation, and innovation while considering the environmental and social impacts of economic activities [155,166]. By addressing these three pillars in an integrated manner, sustainability seeks to create a balanced and harmonious society that respects the limits of our planet's resources and fosters a better quality of life for current and future generations.

Within this context, this literature review section delves into the domains of sustainability, and sustainable products. Through an extensive review of the existing literature, this section aims to explore and critically analyze the interplay of these themes, examining their potential implications in the context of e-commerce and consumer well-being.

### **Sustainable Products**

The importance of sustainable products within the broader concept of sustainability lies in their ability to serve as catalysts for positive change [160]. Sustainable products embody responsible production and consumption practices, helping to transition towards a more sustainable future [156]. By adhering to eco-friendly materials, minimizing waste, and prioritizing ethical sourcing, sustainable products significantly reduce environmental impact throughout their lifecycle [166]. They promote social well-being by upholding fair labor practices, supporting local communities, and ensuring the welfare of workers [163]. Moreover, sustainable products contribute to the long-term viability of the economy by driving innovation, creating job opportunities in green industries, and enhancing market competitiveness [165]. They demonstrate that economic growth and environmental responsibility are not mutually exclusive, but rather interconnected elements that can thrive together.

In terms of environmental conservation, sustainable products prioritize the use of renewable resources and eco-friendly materials, reducing carbon emissions, minimizing waste, and preserving

natural habitats [160,166]. Socially, they enhance the well-being of individuals and communities by ensuring fair wages, safe working conditions, and equitable treatment throughout the supply chain, fostering social equity and empowerment [164,165]. Additionally, sustainable products often support local communities through various initiatives, such as sourcing ingredients locally or reinvesting a portion of profits into community development programs [167]. From an economic perspective, they drive market innovation, attract environmentally conscious consumers, inspire competitors to adopt sustainable practices, create job opportunities in green sectors, and stimulate socially and environmentally responsible economic growth [164]. Overall, sustainable products serve as powerful agents for positive change, addressing environmental, social, and economic challenges for a more sustainable and prosperous future [166].

### **Main Sustainable Product Features**

Having gained an understanding of the classification of sustainable attributes into distinct categories, we proceed to outline a selection of prevalent and pervasive sustainable product features. This compilation serves as a valuable resource for our ongoing research, as it equips us with the essential knowledge to delve deeper into the subject matter. By familiarizing ourselves with these widely recognized sustainable product features, we aim to enhance our understanding of their significance and implications within the broader realm of sustainability.

- **Organic:** The term “organic” refers to a specific attribute of a product, particularly in the context of food and agriculture. Organic products are grown, produced, and processed using methods that prioritize environmental sustainability and the avoidance of synthetic chemicals [168]. Organic farming practices focus on enhancing soil health, biodiversity, and ecological balance, while minimizing the use of synthetic fertilizers, pesticides, antibiotics, and genetically modified organisms (GMOs) [169]. Organic products are cultivated using natural fertilizers like compost or manure, and their pest and disease control measures rely on natural methods such as crop rotation, biological pest control, and the use of beneficial insects. By adhering to these principles, organic products aim to reduce environmental impacts, promote soil fertility, protect natural ecosystems, and offer consumers a healthier and more sustainable choice [170]. Organic certification standards, such as those set by regulatory bodies or third-party organizations, ensure that products labeled as “organic” meet specific criteria and undergo regular inspections to maintain their organic integrity [171,172].
- **Fair trade:** Fair trade is a product feature or certification emphasizing ethical and equitable trading practices, particularly in international trade with producers in developing countries [173]. It ensures that producers receive fair compensation, promotes sustainable livelihoods, and supports community development [173]. Fair trade certification guarantees that the product has been sourced from producers who meet specific social, economic, and environmental standards [174]. By supporting fair trade products, consumers contribute to the empowerment of small-scale farmers, artisans, and workers, addressing social and economic inequalities while promoting sustainable development [175]. Fair trade serves as a significant feature in the realm of sustainable products by fostering fairer trade relationships and promoting the well-being of marginalized producers [176].
- **Recyclable:** The “recyclable” feature is a crucial attribute of sustainable products, as it promotes the efficient use of resources and minimizes waste [177]. Recyclable products can be collected, processed, and transformed into new materials or products, reducing the demand for virgin resources and the environmental impact of extraction [178]. By designing products with recyclability in mind, manufacturers contribute to the circular economy, where materials are kept in use for as long as possible [179]. Recycling also helps reduce landfill waste and the associated environmental and health hazards, making it essential to achieving a more sustainable and resource-efficient society [177].

- Reusable: The “reusable” feature offers an alternative to single-use items that contribute to waste and resource depletion. Reusable products are designed to be used multiple times, reducing the need for constant replacements [180]. By encouraging reuse, these products minimize the environmental impact associated with manufacturing, transportation, and disposal [181]. Reusable products can include items such as shopping bags, water bottles, coffee cups, and food containers, providing consumers with sustainable choices that reduce their reliance on disposable options [180]. By embracing reusable products, individuals can significantly contribute to waste reduction and the conservation of resources, ultimately promoting a more sustainable and environmentally friendly lifestyle [181].
- Renewable energy: The “renewable energy” feature involves harnessing energy from renewable sources that are naturally replenished. By utilizing renewable energy sources like solar, wind, hydro, or geothermal power, sustainable products reduce reliance on fossil fuels and contribute to mitigating climate change [182]. Incorporating renewable energy into product design and production processes helps minimize greenhouse gas emissions and air pollution, promoting cleaner and more sustainable energy systems [183]. Furthermore, the adoption of renewable energy in manufacturing can lead to energy cost savings, increased energy independence, and a more resilient and sustainable economy [182]. By embracing renewable energy, sustainable products contribute to a greener future and support the transition to a low-carbon and sustainable energy system.
- Energy efficient: The “energy efficient” feature is a crucial characteristic of sustainable products as it helps minimize energy consumption and reduce environmental impact [184]. Energy-efficient products are designed to use less energy while maintaining their functionality and performance. By incorporating energy-efficient technologies and design principles, these products contribute to lower energy bills, reduced greenhouse gas emissions, and more sustainable use of resources [185]. Additionally, energy-efficient products often meet or exceed strict energy efficiency standards set by regulatory bodies, ensuring their reliability and effectiveness. By promoting energy efficiency, sustainable products help combat climate change, conserve natural resources, and support a more sustainable and energy-conscious lifestyle [184].
- Sustainable packaging: “Sustainable packaging” focuses on minimizing the environmental impact of product packaging. It involves using materials and design strategies that prioritize resource efficiency, waste reduction, and recyclability [186]. Sustainable packaging aims to minimize the use of non-renewable resources, such as petroleum-based plastics and encourages the use of renewable and biodegradable materials. Additionally, sustainable packaging considers the entire lifecycle of the product, including its production, transportation, use, and end-of-life disposal, striving for minimal environmental impact at each stage [187]. By adopting sustainable packaging practices, products can contribute to reducing plastic pollution, conserving resources, and promoting a more circular and environmentally friendly economy [186].
- Biodegradable/compostable: The “biodegradable/compostable” feature is a significant attribute of sustainable products as it relates to their ability to break down naturally and return to the environment without leaving harmful residues [188]. Biodegradable products can be broken down by microorganisms into natural elements over time, minimizing their impact on ecosystems [189]. Compostable products, on the other hand, are specifically designed to break down in composting systems, resulting in nutrient-rich soil that can be used for agricultural purposes [190]. The use of biodegradable or compostable materials reduces waste and landfill burden, promotes sustainable waste management practices, and supports the transition towards a more circular economy [188]. By choosing biodegradable or compostable products, consumers can make a positive impact on environmental conservation and contribute to a more sustainable future.

- **Low carbon footprint:** The “low carbon footprint” feature is a sustainable product attribute signifying products’ minimal contribution to greenhouse gas emissions and climate change [191]. Products with a low carbon footprint are designed, manufactured, and distributed in a manner that minimizes energy consumption and the release of carbon dioxide and other greenhouse gases. This is achieved through various strategies such as utilizing energy-efficient production processes, sourcing materials locally to reduce transportation emissions, and using renewable energy sources [192]. By reducing their carbon footprint, these products help mitigate climate change, preserve natural resources, and promote a more sustainable and environmentally responsible lifestyle [191]. Choosing products with a low carbon footprint allows consumers to make environmentally conscious choices and contribute to a greener and more sustainable future.
- **Carbon neutral:** The “carbon neutral” feature indicates that the emissions generated throughout products’ lifecycles have been effectively offset or balanced by supporting projects that reduce or capture an equivalent amount of greenhouse gases [193]. Carbon neutrality is achieved by implementing measures to minimize emissions during production, transportation, and disposal, and then offsetting the remaining emissions through activities like reforestation, investing in renewable energy projects, or supporting carbon capture and storage initiatives [194]. By being carbon neutral, these products have a net-zero impact on greenhouse gas emissions, making them environmentally responsible choices [193]. Supporting carbon neutrality in product development helps combat climate change, fosters a more sustainable economy, and encourages a shift towards a low-carbon future [194]. Choosing carbon-neutral products allows consumers to make a positive environmental impact and contribute to global climate action [194].
- **Ethically sourced:** The “ethically sourced” feature focuses on ensuring that the raw materials used in products’ production are obtained in a socially and environmentally responsible manner [195]. Ethically sourced products prioritize fair trade, respect for human rights, and adherence to labor standards throughout the supply chain [196]. This includes fair wages, safe working conditions, and the prohibition of child labor or exploitation [196]. By choosing ethically sourced products, consumers can support companies that prioritize social equity, worker well-being, and environmental stewardship [195]. Ethical sourcing promotes transparency, accountability, and sustainable practices, making it an essential aspect of sustainable product selection [197].
- **Water conservation:** The “water conservation” feature aims to minimize water usage and promote responsible water management throughout the product lifecycle [198]. Products designed with water conservation in mind employ strategies such as efficient water use, water-saving technologies, and responsible manufacturing processes that reduce water waste and pollution [199]. By minimizing water consumption, these products help alleviate water scarcity, preserve freshwater ecosystems, and protect this vital resource for future generations. Additionally, water conservation in product design encourages consumer awareness and responsible water usage, fostering a more sustainable and environmentally conscious mindset [199]. Choosing products with a water conservation feature allows individuals to contribute to water sustainability and promote a more water-efficient society.
- **Non-toxic:** The “non-toxic” feature ensures that products are free from harmful substances and pose minimal risks to human health and the environment [200]. Non-toxic products are designed and manufactured using materials that do not contain toxic chemicals or pollutants, including heavy metals, harmful solvents, or volatile organic compounds (VOCs) [201]. By prioritizing non-toxic ingredients and materials, these products minimize the potential for adverse health effects, allergies, or pollution during production, use, and disposal [200]. Choosing non-toxic products promotes a safer living environment, reduces the burden on ecosystems, and supports sustainable consumption practices [200]. Furthermore, non-toxic

products often undergo rigorous testing and certification processes to provide consumers with assurance regarding their safety and environmental impact [201].

- **Cruelty-free:** The “cruelty-free” feature indicates that products have been developed without any involvement or harm to animals throughout their production process [202]. Cruelty-free products are not tested on animals and do not contain ingredients derived from animals or animal by-products [203]. By choosing cruelty-free products, consumers can support ethical and compassionate practices in the beauty, personal care, and household industries [203]. This feature promotes the well-being and rights of animals, encourages the development of alternative testing methods, and contributes to the overall shift towards more ethical and sustainable choices [202]. Additionally, cruelty-free products often carry certifications or labels from reputable organizations, providing transparency and assurance to consumers seeking animal-friendly options [203].
- **Upcycled:** The “upcycled” feature is a significant attribute of sustainable products, indicating that they have been created by repurposing or transforming waste materials into new and valuable items [204]. Upcycling is a process that diverts materials from landfill and reduces the need for new resource extraction. By upcycling, products can give new life to materials that would otherwise be discarded, contributing to waste reduction and resource conservation [204,205]. Upcycled products showcase creativity and innovation in design, highlighting the potential for transforming waste into unique and functional items [205]. Choosing upcycled products promotes a circular economy mindset and supports sustainable production practices that minimize waste and maximize the value of existing resources.
- **Locally produced:** The “locally produced” feature emphasizes products’ origin from nearby sources, typically within a specific geographic region [206]. Locally produced products support local economies, reduce carbon emissions associated with long-distance transportation, and promote community resilience [207]. By choosing locally produced items, consumers can contribute to the vitality of their local communities and support local artisans, farmers, and businesses. Furthermore, local production often involves shorter supply chains, allowing for better transparency, traceability, and a closer connection between consumers and producers [207]. This feature encourages sustainable consumption by fostering a sense of place, reducing environmental impact, and strengthening local economies [206].
- **Social impact:** The “social impact” feature highlights products’ positive influence on society and communities. Products with a focus on social impact contribute to social welfare, equality, and empowerment [208]. They may support fair trade practices, provide employment opportunities for marginalized groups, or invest a portion of their profits in social initiatives [173,174]. By choosing products with a social impact, consumers can align their purchasing power with their values and actively support positive change in society [209]. These products often go beyond profit-driven motives and strive to create a lasting, positive difference in the lives of individuals and communities, making them an integral part of the sustainable product landscape [209].
- **Low/zero VOC:** The “low/zero VOC” indicates that products contain minimal or no volatile organic compounds (VOCs) [210]. VOCs are chemicals that can be released into the air and contribute to indoor and outdoor air pollution, posing risks to human health and the environment [210]. Low/zero VOC products prioritize the use of non-toxic or low-emission materials, such as paints, adhesives, and cleaning products [201]. By choosing low/zero VOC products, consumers can create healthier indoor environments, reduce their exposure to harmful chemicals, and contribute to improved air quality [210,211]. This feature aligns with sustainable practices that prioritize human well-being and environmental stewardship [197].

### 2.9.3. Materials and Methods

In **Section 2.9.2**, the concept and necessity of sustainability were analyzed, as well as the characteristics that products must possess in order to be considered sustainable. It was also found that

sustainability is not solely about the material composition of a product and its environmental footprint, but social and economic factors can also influence it. In **Sections 2.2 to 2.5**, the GPT Models and NLP were discussed as emerging AI technologies. The concept of pre-training and fine-tuning in GPT models was explained, along with how prompts are used to guide the generation of text, and how GPT models contribute to advancing NLP capabilities.

### **Research Objectives**

The primary objective of this study is to propose a novel approach for automating the identification of sustainable product features using GPT-powered software. The proposed software will leverage NLP techniques to classify products into distinct sustainability categories based on their descriptions, titles, and product specifications.

#### *Expected Outcomes and Benefits*

By achieving this research objective, we anticipate several significant outcomes and benefits:

- **Enhanced Consumer Empowerment:** The GPT-powered software will empower consumers to make more informed and sustainable purchasing decisions effortlessly. By automatically extracting sustainability indicators from product descriptions, titles, and product specifications, consumers will have access to crucial environmental and ethical information at their fingertips, enabling them to align their purchasing choices with their sustainability values.
- **Time and Resource Savings:** The automated sustainability assessment through the GPT-powered software will streamline the evaluation process for consumers and businesses alike. The software will eliminate the need for time-consuming and labor-intensive manual assessments, saving valuable resources and allowing businesses to allocate their efforts to other sustainability initiatives.
- **Environmentally Conscious Business Practices:** The proposed approach will encourage businesses to showcase and prioritize their environmentally friendly offerings. By providing automated sustainability classifications for products, the software will motivate businesses to adopt sustainable practices, ultimately contributing to a more eco-conscious market.

#### *Significance and Contribution to Sustainability Research*

This research is highly significant in addressing the existing challenges in sustainability assessment. The increasing global focus on sustainability necessitates efficient and accurate methods for evaluating product eco-friendliness. The proposed GPT-powered software represents an innovative fusion of advanced language models with product classification, bridging the gap between NLP and sustainability evaluation.

The integration of GPT technology in sustainability assessment not only ensures more comprehensive and objective analysis but also extends the applicability of NLP to environmental and social domains. By combining language-understanding capabilities with ML techniques, the software can capture subtle nuances and context-specific sustainable features that might be challenging to identify using traditional methods.

The contribution of this research lies in its potential to revolutionize the way sustainable product features are identified and communicated. By automating sustainability assessment, the proposed approach paves the way for a more sustainable and eco-conscious future. It facilitates a shift towards greener consumption patterns, which can significantly reduce the environmental impact of products and promote circular economy practices.

Overall, the proposed approach holds promising implications for both consumers and businesses, fostering a more sustainable marketplace and supporting efforts towards achieving global sustainability goals. It represents a valuable addition to the field of sustainability research, offering an innovative and efficient solution to address the pressing challenges of sustainability assessment across various industries.

## Methodology

To validate the efficacy of the software and the suitability of the GPT model for assessing and identifying sustainable features in products, a specific methodology was employed. A representative sample of products from three well-established marketplaces was gathered, and the GPT model was utilized to evaluate and identify the sustainable features of each product. The subsequent sections detail both the data collection and preprocessing process, as well as the model evaluation based on the responses obtained.

### *Data Collection and Preprocessing*

Ensuring data quality and relevance is crucial to maintain the integrity and reliability of the research findings. In this study, the following data-preprocessing steps were undertaken to achieve this:

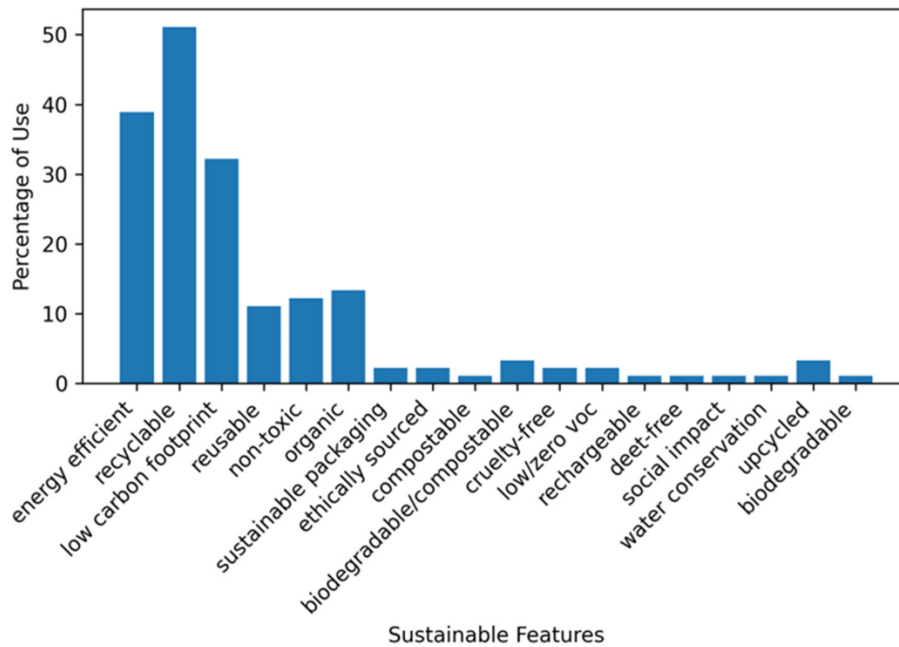
- **Data Collection:** For the purpose of this research, we utilized the most well-known marketplaces, namely Amazon, eBay, and AliExpress, as reference points. From each of these marketplaces, we collected a total of 30 randomly selected products, spanning across 6 distinct product categories. Care was taken to cover various industries and product categories to ensure the software's and GPT model's applicability across different domains.
- **Data Cleaning:** Raw text data often contain noise, irrelevant information, or typographical errors. Data cleaning involved removing special characters, punctuation, and irrelevant tags or metadata. Additionally, any duplicated or redundant entries were eliminated to enhance data quality.
- **Text Normalization:** Text normalization is crucial to standardize the data, ensuring consistency in word representations. This step involves converting all text to lowercase, handling contractions, and applying stemming or lemmatization techniques to reduce words to their base form.
- **Tokenization:** The process of tokenization breaks down the textual data into individual tokens (words or subwords). Tokenization is a critical step for language models like GPT, as it allows the software to process and understand text at a granular level.
- **Feature Extraction:** Relevant features related to sustainability, such as "eco-friendly materials," "energy efficiency," "recyclability," and "ethical sourcing," will be identified and extracted from the preprocessed text. This step is vital to building a robust classification model that can differentiate products based on their sustainability characteristics.
- **Software assessments:** With the GPT-powered Chrome app activated, we accessed the webpage of each product, and the application was tasked with extracting the product title, description, and available product specifications, as applicable. Subsequently, the collected data were transmitted to a remote server, which established communication with the appropriate GPT model, soliciting responses concerning the sustainability features of the product based on its title and description. The resulting outcomes were stored in a CSV file to facilitate further analysis. A comprehensive description of this process can be found in **Software Implementation and System Design** section. The sample products and software's assessments are readily available in CSV format on GitHub repository [212].

By rigorously implementing these data preprocessing steps, we aim to guarantee the quality and relevance of the dataset used to test the GPT-powered software. These steps ensure that the software can effectively and accurately identify sustainable product features across various domains, providing reliable and actionable insights for consumers and businesses alike.

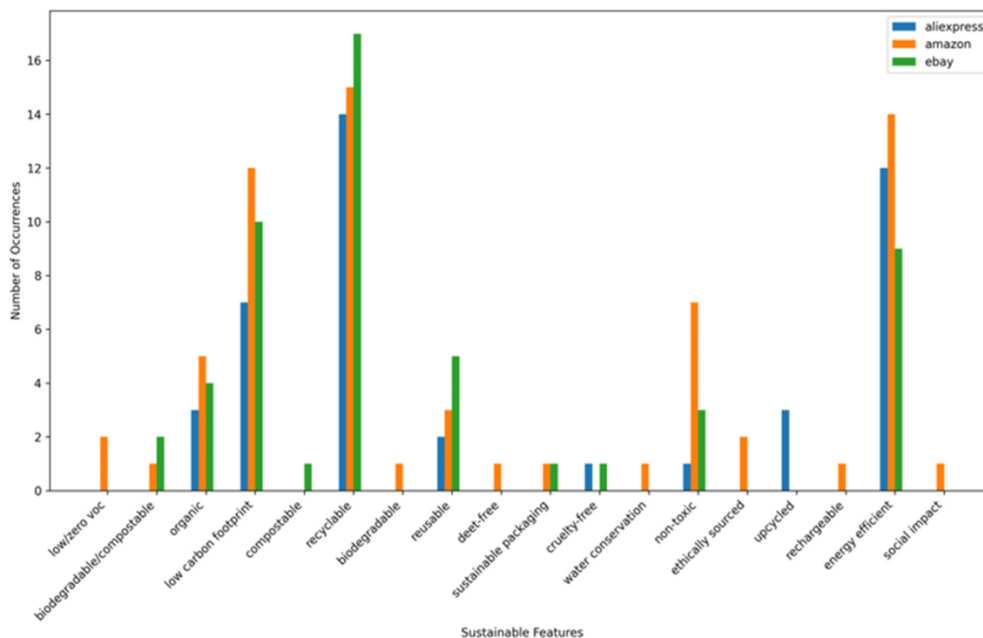
Upon data collection and processing, the CSV file was imported into a Python script responsible for populating a data frame (df) to facilitate the extraction of meaningful plots. For this purpose, two Python libraries, namely `pandas` and `matplotlib`, were utilized. `Pandas` is a robust Python library extensively employed in data manipulation and analysis tasks, offering essential data structures such as data frames that facilitate the effective management of substantial datasets [213]. On

the other hand, `matplotlib`, a widely adopted Python library, specializes in data visualization endeavors, providing a versatile selection of plotting techniques to generate diverse graph types and visual depictions derived from data stored in arrays or data frames [214].

In **Figure 10**, the percentage of utilization of sustainable features among the products in our dataset is presented, while in **Figure 11**, the frequency of sustainable feature occurrences categorized by marketplace is illustrated.



**Figure 10.** Percentage of Use of Sustainable Features Across Products.



**Figure 11.** Number of Occurrences of Sustainable Features by Marketplace.

### Model Evaluation

As stated in **Section 2.4**, the GPT model utilized in this research is a pre-trained model that has undergone extensive training on diverse textual datasets from various sources. This pre-training process exposed the GPT model to vast amounts of text, enabling it to learn language patterns, syntactic

structures, and contextual relationships. The pre-training phase aimed to establish a robust language understanding foundation, which is crucial for subsequent fine-tuning on specific tasks.

For the purposes of this research, the GPT pre-trained model was utilized without undergoing fine-tuning. This decision was based primarily on the strong language and contextual capabilities of the pre-trained model, as well as its training on diverse sources of textual data. Another reason was to assess the quality of responses generated by the pre-trained model for this specific task.

To evaluate the performance of the GPT-powered software in accurately identifying sustainable attributes, a manual inspection of the 90 sampled products was conducted. Additionally, the evaluation process incorporated a comprehensive qualitative analysis of the results. This qualitative assessment sought to identify potential patterns, limitations, or challenges encountered by the GPT-powered software in sustainable feature identification. The objective of this analysis was to gain valuable insights into the model's decision-making process and to pinpoint areas for potential improvements in the proposed approach.

#### **2.9.4. Experimental Setup and Results**

##### **User Interaction**

The initial idea was to create a desktop or mobile app that would allow consumers to receive an answer about whether a product has sustainable features or not, either by pasting a link or a product description. However, adding another mobile app to the thousands already existing on our mobile devices, which are likely to never be used, seemed counterproductive. The rationale from the beginning was to facilitate the user making a more sustainable purchasing decision without burdening them in any way. For this reason, we chose to develop a Google Chrome app that, once installed, would function autonomously without any user intervention.

The developed software facilitates user interaction with an e-commerce or marketplace platform. Upon accessing a product page, the accompanying Chrome application assumes the responsibility of extracting the product's title, description, and features. Through communication with a remote server, employing a GPT model, the gathered information undergoes analysis to ascertain the presence or absence of sustainable attributes within the product. Consequently, the system provides the user with a response indicating whether the currently viewed product exhibits sustainable characteristics. In summary, the user's act of opening a product page prompts the GPT-powered Chrome application to deliver information essential for making informed sustainable purchasing decisions.

##### **Software Implementation and System Design**

The implementation of the software involved the utilization of various technologies and programming languages. JavaScript, JQuery, and Ajax with Manifest version 3.0 were employed for creating the Chrome APP/Extension [214]. On the server side, Python was utilized along with the Flask framework and relevant libraries. For the implementation of the GPT component, Python was responsible for invoking the OpenAI API [151], retrieving the response, and returning it to the Chrome APP, where JavaScript would handle the printing of the results to the user.

A detailed description is provided below, outlining the steps undertaken by the Chrome app to identify and make a decision regarding the sustainable attributes of a product.

- The user proceeds with the installation of the Chrome app on the Google Chrome browser, utilizing the extensions area in developer mode;
- The user navigates through one of the prominent marketplaces/e-commerces, such as Amazon, and arrives at a product page;
- The background-running Chrome app detects the user's presence on a product page and utilizes JavaScript and JQuery to extract the title, description, and characteristics of the product;
- Through Ajax, the Chrome app initiates an asynchronous POST request to a remote server;

- The script is hosted on PythonAnywhere [215], a project associated with Anaconda [216], renowned among both data scientists and other users;
- The server employs Flask [217], a micro Python-based framework, to handle the incoming POST request from the Chrome app;
- Python is responsible for loading the OpenAI library and generating an appropriate prompt containing the title and description of the product observed by the user;
- Upon receiving the response from the OpenAI API, Python undertakes the processing of the response data and formats them in a manner suitable for the Chrome app to interpret. These formatted data are then returned to the Chrome APP;
- Depending on the response, the Chrome app asynchronously displays a badge on the page being viewed by the user. On mouse hover, a popup message appears, presenting text that elucidates the sustainable attributes identified by the GPT for the product. In the scenario where no sustainable attributes are detected, a distinct badge is printed on the page, which, upon hover, notifies the user that no sustainable attributes were found for the specific product.

**Figure 12** depicts the prompt employed to elicit the desired response from the GPT model and OpenAI API. The placeholder `{product_title}` is automatically replaced with the product title, `{product_description}`, which represents the product description, and `{sustainable_attributes_to_string}` corresponds to a stringified array containing all the sustainable attributes analyzed in **Main Sustainable Product Features** section.

```
prompt = f"I have a product titled '{product_title}' and product
description '{product_description}'. Match the product with
sustainable attributes and explain why. Sustainable attributes:
{sustainable_attributes_to_string}. If the product can't match any
of the sustainable attributes, return False.\n Else, return a
dictionary like this: {'attribute': 'Explanation'}"
```

**Figure 12.** Prompt to obtain the appropriate response from the GPT model.

For the purposes of this study, the `text-DaVinci-003` model, which belongs to the GPT-3 model family, was employed [218].

The `text-DaVinci-003` model incorporates several enhancements, including the ability to generate higher-quality writing, thereby facilitating the delivery of clearer, more engaging, and more compelling content through applications. Moreover, it demonstrates improved proficiency in handling complex instructions, enabling users to explore more creative ways to leverage its capabilities. Additionally, the model exhibits enhanced competence in generating longer-form content, expanding its utility to encompass tasks that were previously deemed challenging [219].

Undoubtedly, the software can be readily upgraded in the future to utilize the GPT-4 model, which is presently undergoing limited beta testing and is exclusively accessible to a select group of authorized individuals [220].

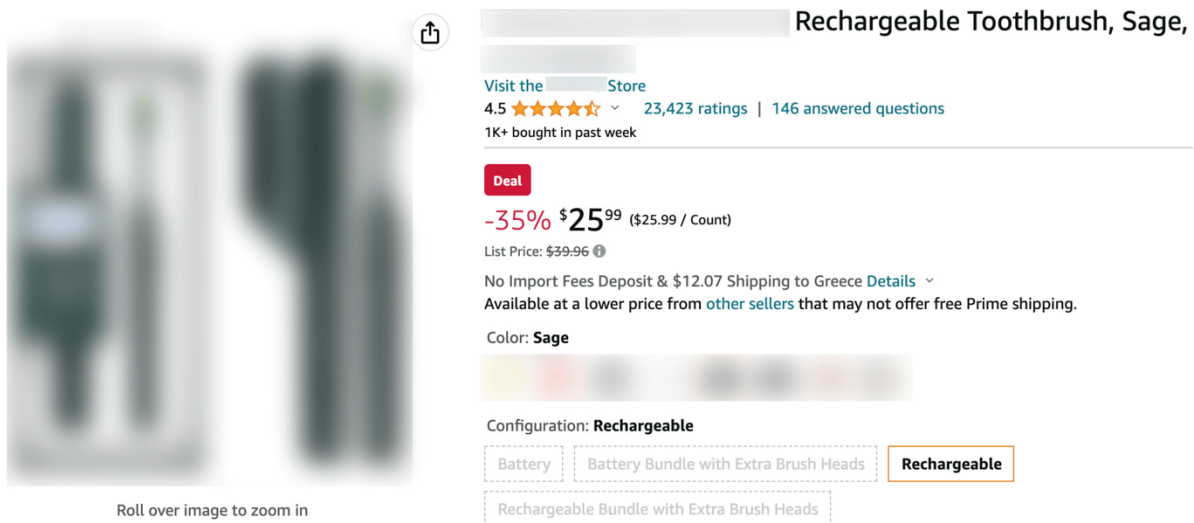
The source code of the Chrome APP, both for the client-side and server-side components, is made available as open-source on GitHub [212].

### *Case Studies and Examples*

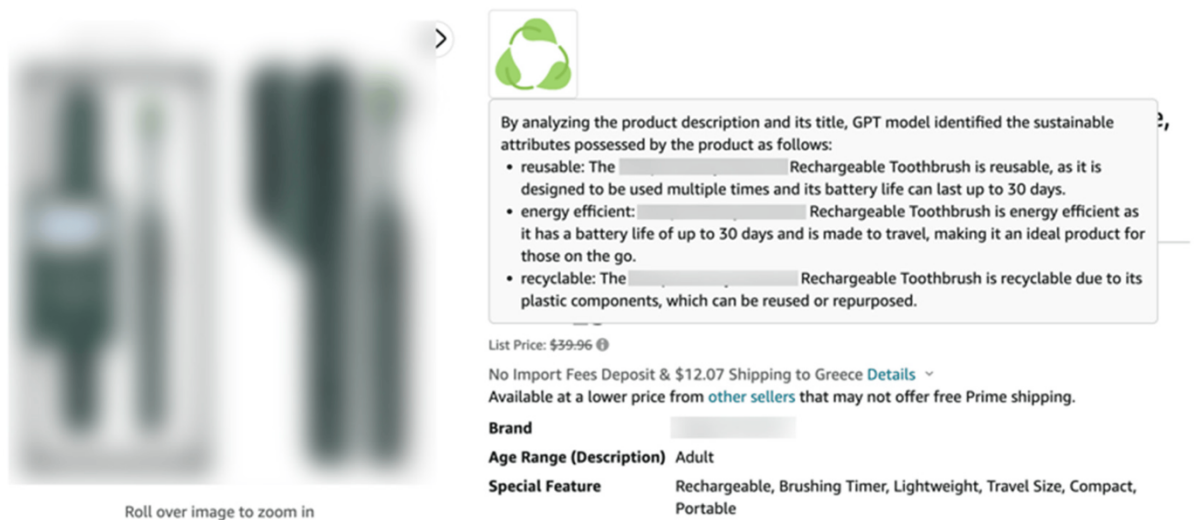
This section presents two case studies pertaining to the functioning of the software when a user navigates a product page with the GPT-powered Chrome app enabled. Subsequently, an associated badge is displayed to the user following the evaluation of the sustainable features of the product. Upon hovering the mouse over the badge, the user receives a textual representation of all the sustainable features that the GPT model identified within the specific product.

### Case Study 1

**Figure 13** presents a visual representation of a product page in its initial state before receiving any response from the Chrome app. In **Figure 14**, the same product page is depicted after the response generated by the Chrome app has been incorporated. Notably, a green badge has appeared above the title, which, upon hovering, displays the GPT model's responses concerning the sustainable features of the product.



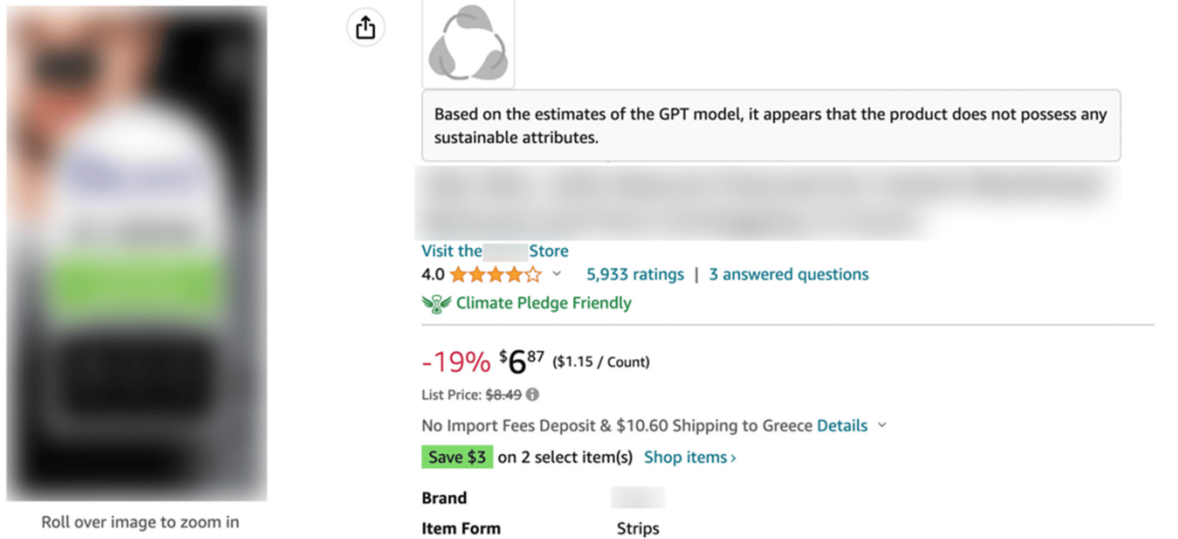
**Figure 13.** Screenshot of a product page before receiving the response from the Chrome app and GPT model.



**Figure 14.** The same product page displaying modifications after the response generated by the Chrome app and GPT model.

### Case Study 2

**Figure 15** displays a non-sustainable product along with the corresponding badge that appears to the user following the evaluation performed by the GPT model.



**Figure 15.** The non-sustainable product page displaying modifications after the response generated by the Chrome app and GPT model.

## Results

The assessment of a domain-specific model typically necessitates the use of domain-specific evaluation metrics. In the domain of sustainability, coupled with the GPT model, there are currently no available evaluation metrics.

To confirm the inadequacy of general evaluation metrics in appraising a sustainable evaluation software, as presented in this study, the BLEU (Bilingual Evaluation Understudy) [221] metric was experimentally employed. Loading the CSV file into the Jupyter Notebook and comparing the product titles and descriptions with the responses generated by the GPT model yielded discouraging results, with a BLEU score of 0.07259360665469289 on average. This score suggests that the model-generated text bears some resemblance to the reference data, but the degree of similarity is relatively low. A higher BLEU score, closer to 1, would indicate a more favorable alignment between the model's outputs and the reference data.

Following the methodology outlined in Methodology Section, human evaluation was conducted to scrutinize each line of the CSV in order to ascertain whether the responses from the GPT-powered software adequately align with the hypotheses presented in the Introduction (**Section 2.9.1**).

### *Hypothesis 1*

- **Research Objective:** The study seeks to evaluate the accuracy of GPT model responses.
- **Null Hypothesis (H0):** There is no significant accuracy in GPT model responses when appropriate prompts are utilized.
- **Alternative Hypothesis (H1):** The GPT model produces accurate responses when appropriate prompts are employed.

The investigation involved the use of multiple prompts in an iterative manner until the desired and appropriate response was achieved from the GPT model. Throughout this process, the generated responses consistently maintained relevance to the topic and demonstrated precision. This observation indicates that tailoring prompts to elicit specific information from the model can effectively enhance the accuracy and pertinence of its outputs. As a result, the null hypothesis H0 can be rejected, leading to the conclusion that with meticulous prompt selection and refinement, the GPT model can serve as a valuable tool for providing accurate and pertinent information across various contexts.

### *Hypothesis 2*

- **Research Objective:** The study aims to examine the format of responses generated by the GPT model.
- **Null Hypothesis (H0):** The responses produced by the GPT model do not conform to the appropriate format.
- **Alternative Hypothesis (H1):** The format of GPT model responses is accurate and aligns with the explicitly defined structure in the prompt.

In the course of the research, multiple prompts were utilized to prompt the GPT model to generate responses in a dictionary format suitable for retrieval in the Chrome app via an Ajax request. However, the returned dictionaries presented multiple choices, leading to instances of truncated responses due to the imposed constraint on the maximum tokens per request by OpenAI. Addressing this issue programmatically involved merging the choices and eliminating duplicates, but this did not provide sufficient evidence to confirm the alternative hypothesis (H1).

As a result, the null hypothesis cannot be rejected, and the encountered limitation in response format indicates the need for further investigation and refinement to achieve the desired response structure. Future research efforts should focus on developing more effective methods to ensure that the GPT model consistently delivers complete and unambiguous dictionary-style responses. By achieving this objective, the usability and applicability of the GPT model within the context of the Chrome app can be significantly enhanced.

### *Hypothesis 3*

- **Research Objective:** The study aims to assess the effectiveness of the GPT model in identifying sustainable features based on product titles and descriptions.
- **Null Hypothesis (H0):** The GPT model is unable to identify sustainable features of a product when given its title and description.
- **Alternative Hypothesis (H1):** The GPT model is capable of evaluating the sustainable features of a product when provided with its title and description.

Following a comprehensive manual evaluation of 90 products within the sample, the responses generated by the GPT model regarding the sustainable features of these products demonstrated a high degree of accuracy. The model exhibited proficiency in recognizing sustainable features through both the identification of relevant keywords and the comprehension of contextual cues present in the product titles and descriptions. As a result, the null hypothesis H0 is rejected, indicating that the GPT model possesses the capability to effectively assess and identify sustainability-related attributes of products. This finding underscores the potential of the GPT model as a valuable tool for sustainable product analysis and evaluation.

### *Hypothesis 4*

- **Research Objective:** The study aims to evaluate the effectiveness of the GPT model in identifying product sustainable features, particularly when product titles and descriptions are less descriptive.
- **Null Hypothesis (H0):** The GPT model is incapable of identifying product sustainable features when confronted with limited or less descriptive product titles and descriptions.
- **Alternative Hypothesis (H1):** The GPT model can successfully retrieve sustainable features of a product even when the provided title and description lack sufficient descriptive information.

Upon conducting a meticulous manual evaluation, notable observations were made, revealing the GPT model's ability to delve into its own training data and retrieve pertinent information about a product, even in scenarios where the product titles and descriptions were inadequately descriptive. Despite encountering limitations in the available details, the model skillfully employed its pre-trained

knowledge to generate accurate responses pertaining to the sustainable features of the products. Consequently, the null hypothesis H0 can be confidently rejected. The GPT model's remarkable adaptability and capacity to complement incomplete information with acquired knowledge from its training data underscore its robustness and versatility in identifying sustainable attributes, regardless of the quality of input descriptions. As a result, the GPT model emerges as a promising tool for effectively assessing sustainability in products with varying levels of descriptive information.

### **2.9.5. Discussion**

In previous sections, an in-depth examination was conducted to elucidate the methodology and tools employed during the software implementation and evaluation process, as well as to address the limitations and challenges encountered. In this section, the authors offer their contemplations regarding potential future enhancements of the software. Additionally, they delve into the ethical considerations surrounding the application of the GPT model in sustainability assessment tasks.

#### **Application of Findings and Future Research Directions**

Commencing with the software, it aspires to serve as a solution with the aim of raising awareness among the broader consumer base and encouraging potentially more sustainable purchasing decisions that may positively impact the environment and socioeconomic aspects of life. Concurrently, it can assist product producers in crafting more comprehensive descriptions that accurately delineate the sustainable attributes of their products. The issue of a lack of a sustainable mindset should not solely rest upon consumers. The issue originates from the lack of appropriate information available to consumers. At the consumer level, if individuals were aware that a product is more sustainable than another, without exerting any effort, they would undoubtedly choose the former over a non-sustainable alternative. Furthermore, e-commerce platforms and marketplaces should also assume a corresponding role, either by developing their own tools to inform consumers about the sustainability features of products or by incorporating the automated open-source solution provided by this study simply by integrating the Chrome app's JavaScript into their underlying source code.

From the perspective of GPT and NLP, their integration into emerging technologies is inevitable, and they undeniably have the potential to serve as the cornerstone of an AI technological revolution. However, it would be prudent for the OpenAI API, in particular, to facilitate development by providing documentation not only on how to communicate with the API but also on how to handle the numerous choices that often differ or are truncated due to token limitations. For instance, the "create" method within the completion class could potentially be enhanced to accept parameters defining the expected format of the output, such as JSON encoded, text, or integer representations. Additionally, since the creation of a suitable prompt is highly significant, the API itself should assist developers in generating appropriate prompts through a corresponding tool. This would aid the GPT model in understanding the intention of the prompt and providing the precise responses that developers expect. The ChatGPT, currently based on GPT 3.5/GPT 4, could be leveraged for generating prompts. A corresponding graphical user interface (GUI) could be integrated into the user panel, enabling developers to input the expected outcomes and their respective formats. The GPT model would then respond with the appropriate prompt, assisting the developers in obtaining the desired responses.

As explicated in the section concerning model evaluation, the GPT pre-trained model was employed in this research without undergoing fine-tuning, aiming to assess its existing capabilities in detecting sustainable product features. Subsequent investigations may consider fine-tuning the GPT model by training it with additional knowledge specifically related to sustainability. However, such fine-tuning would necessitate a substantial and well-balanced dataset to accurately train the GPT model to perform sustainability tasks more effectively.

## Software Limitations and Challenges

Undoubtedly, no software is immune to problems and challenges, which can be addressed and are presented below:

- Prompt-related issues: To achieve the desired outcome described in **Software Implementation and System Design** section, numerous attempts were made to find the appropriate prompt that would guide the GPT model in understanding the required actions. For instance, despite prompting the model to evaluate the product based on specific attributes, it would still generate its own assessments for attributes not provided as options. This problem was resolved by refining the prompt.
- Efforts were also devoted to handling the responses returned by the OpenAI API. After each response, the API provided multiple choices, with each choice containing fragmented answers due to token limitations. By refining the prompt and providing the exact expected format of the result, the answers became more accurate. However, the presence of multiple choices continued to complicate the extraction of results.
- At present, the software is fully functional with the Amazon marketplace, which has well-structured source code using CSS classes and IDs. In contrast, in similar attempts on eBay, it proved almost impossible to extract product descriptions using JavaScript, as the eBay rich description operates on iframe technology. This issue can be addressed by accessing the product information through the official APIs provided by the marketplaces instead of relying on JavaScript-based extraction.

## Ethical Considerations

Ethical considerations surrounding the utilization of the GPT model for identifying sustainable features of a product based solely on its title and description are of paramount importance. While the model's ability to analyze textual data and provide insights can be valuable in promoting sustainable practices, certain ethical concerns must be acknowledged. One major concern is the potential for biased or inaccurate assessments due to the limitations of context and comprehensiveness in the provided information. Relying solely on the title and description may overlook critical details, leading to incomplete or misleading conclusions. Moreover, there is a risk of perpetuating greenwashing practices, where products are falsely labeled as sustainable, leading to consumer deception. To address these ethical challenges, researchers and developers must ensure transparency in the methodology, actively acknowledge and disclose the model's limitations, and consider incorporating supplementary materials sources to enhance the accuracy and reliability of the sustainability assessments. Additionally, implementing robust validation mechanisms and engaging with domain experts can help mitigate potential ethical pitfalls and foster responsible and meaningful use of the GPT model for sustainability evaluations.

## Chrome Extensions and APPs in Similar Contexts

In the market, various chrome extensions and mobile applications are available, mirroring the proposal presented in this research. These applications, like our GPT-powered extension, contribute to the promotion of sustainability and incentivize consumers to make more informed and sustainable decisions. For instance, the Ethical Shopper chrome extension informs consumers about the sustainability practices adopted by different brands. The Ecowiser chrome extension aids consumers in selecting more sustainable products from marketplaces like Amazon, by identifying corresponding keywords in the recommended products. Additionally, the Good On You Android app serves as a valuable resource for sustainability ratings in the realm of fashion.

The achievement of sustainability is further attainable with the availability of appropriate tools and applications. By leveraging the capabilities of GPT-powered technologies, along with other innovative approaches, consumers are better equipped to align their choices with sustainability principles, fostering a more eco-conscious future.

### **2.9.6. Section Conclusion**

In conclusion, this study highlights the growing importance of sustainability in today's industries and the increasing consumer demand for sustainable products. It introduces a novel approach, utilizing GPT-powered software to automatically identify sustainable product features from product descriptions, titles, and product specifications. By leveraging NLP and ML, the software successfully classifies products into different sustainability categories and extracts key indicators of sustainability. This automated process eliminates the need for manual assessment and enhances the evaluation of product sustainability. The software empowers consumers to make more sustainable purchasing decisions and enables businesses to showcase their environmentally friendly offerings. The experimental results demonstrate the effectiveness and accuracy of the GPT model in the field of sustainability assessment, rendering it a valuable tool for promoting sustainability and paving the way for a more eco-conscious future. Prospective investigations hold the promise of further narrowing the gap between AI models and sustainability, where the latter will be meticulously fine-tuned for specific sustainability objectives. The realization of such advancements bears the potential to augment and optimize human existence, thereby fostering heightened eco-consciousness in decision-making processes.

## ***2.10. Churn Modeling in E-commerce: Leveraging the Power of Logistic Regression and Neural Networks***

### **2.10.1. Section Summary and Significance of the Study**

In the dynamic realm of e-commerce, understanding and predicting customer churn has become imperative for businesses aiming to sustain growth and profitability [222]. The challenge of churn, where customers disengage from a brand or platform, poses significant hurdles for e-commerce enterprises. Churn modeling emerges as a critical facet of this effort, offering businesses insights into the factors influencing customer attrition [223]. Recognizing the imperative for proactive strategies in mitigating churn, this study aims to investigate the dual-pronged approach of harnessing both Logistic Regression and Neural Network models to elevate the accuracy and effectiveness of churn prediction within the e-commerce domain.

Logistic Regression, a well-established statistical technique, proves to be a robust tool for scrutinizing the relationships between various predictor variables and the likelihood of customer churn [224]. Its interpretability and simplicity make it an ideal choice for comprehending the drivers behind customer attrition. Conversely, Neural Networks, with their capacity to capture intricate non-linear patterns, provide a more profound and nuanced understanding of customer behavior [225]. By harnessing the strengths of these methodologies, the objective is to leverage their complementary features and train them specifically for a specialized churn prediction task, ultimately leading to more precise predictions of customer churn in e-commerce domain.

Through an exploration of a specific real-world dataset, this study aims to train the Logistic Regression and Neural Network models for churn modeling, compare their performance, and extract valuable insights. Synthesizing insights derived from these models allows businesses not only to identify customers at risk of churning but also to formulate targeted strategies for retention and re-engagement.

### **2.10.2. Literature Review**

In the dynamic realm of E-commerce, the study of churn modeling takes center stage as businesses grapple with the pervasive challenge of customer attrition. Providing a comprehensive overview of customer churn within the E-commerce industry is crucial for illuminating the complexities associated with customer retention in this rapidly evolving digital landscape. Customer churn, defined as the rate at which customers disengage from a platform, holds immense significance

for businesses seeking sustained growth and profitability [226]. Effectively addressing and understanding churn becomes imperative not only for maintaining a loyal customer base but also for optimizing operational efficiency. The intricacies of customer behavior and the myriad factors contributing to churn necessitate a strategic and data-driven approach [227]. This is where the importance of predictive modeling in E-commerce becomes evident.

Leveraging advanced analytical techniques, such as logistic regression and neural networks, empowers businesses to anticipate and proactively manage customer churn. The subsequent exploration of predictive modeling in the context of E-commerce underscores its pivotal role in developing effective strategies for customer retention, fostering business resilience in the face of competitive markets and evolving consumer preferences.

### **Historical Perspectives on Churn Modeling**

Historically, churn modeling within E-commerce has witnessed significant evolution and innovation. In its early stages, methodologies focused on rudimentary models, relying on basic statistical analyses to understand customer behavior. As the E-commerce landscape matured, so did the sophistication of churn modeling techniques [228]. Key advancements, including the incorporation of ML algorithms and predictive analytics, reshaped the landscape, allowing for more nuanced and accurate predictions. Recognizing the historical journey of churn modeling provides valuable insights into challenges overcome, lessons learned, and the continuous refinement of methodologies in the contemporary understanding and application of customer churn.

### **Frameworks for Customer Retention**

Theoretical frameworks related to customer retention form the cornerstone for effective strategies within the E-commerce domain. These frameworks provide a conceptual foundation, guiding businesses in devising practical approaches to retain customers in a highly competitive environment. In the application of customer retention strategies in E-commerce, businesses draw upon a rich tapestry of theoretical models, ranging from traditional loyalty theories to contemporary behavioral economics [229]. Strategies encompass personalized marketing, loyalty programs, and enhanced customer experiences. Analyzing these insights enables E-commerce practitioners to tailor their approaches, fostering a deeper understanding of customer needs and preferences.

### **Logistic Regression in Churn Modeling**

Introducing logistic regression as a predictive modeling technique in the realm of E-commerce churn modeling signifies a methodological cornerstone. Logistic regression, a statistical approach well-suited for binary outcome predictions, has gained prominence for its simplicity and interpretability. In the context of E-commerce churn modeling, numerous studies have harnessed the power of logistic regression, showcasing its effectiveness in predicting customer attrition [224]. Its strength lies in its ability to provide transparent insights into the relationship between independent variables and the probability of churn, making it accessible and practical for businesses seeking actionable intelligence.

### **Neural Networks in Churn Modeling**

Delving into the role of neural networks in churn modeling within the E-commerce landscape necessitates an exploration of their overarching capabilities in predictive analytics. Neural networks, inspired by the human brain's interconnected structure, excel in discerning intricate patterns and relationships within vast datasets. In the context of E-commerce churn prediction, studies leveraging neural networks have showcased their ability to uncover nuanced customer behavior, offering insights beyond the reach of traditional modeling techniques [222].

### **Customer Behavior and Segmentation in E-commerce**

A thorough exploration of customer behavior and segmentation within the domain of E-commerce serves as a foundational element in the pursuit of effective churn modeling. The literature

review in this area encompasses a diverse range of studies focused on understanding the intricacies of customer behavior [230]. This examination is crucial for discerning patterns, preferences, and triggers that influence customer engagement and retention. Equally significant is the emphasis on segmentation, allowing businesses to categorize customers based on shared characteristics, needs, or behaviors. Understanding customer behavior and implementing segmentation strategies becomes pivotal in the context of churn modeling, allowing for the identification of high-risk segments and tailoring retention strategies accordingly.

### **Business Intelligence for Churn Management**

The integration of business intelligence tools into churn management represents a strategic imperative for businesses navigating the complexities of customer retention in the E-commerce sphere. The literature on this integration delves into the diverse array of business intelligence tools available and their applications in deciphering and mitigating customer churn [231]. Case studies illuminate successful instances where business intelligence has been instrumental in identifying patterns, trends, and customer insights crucial for effective churn management strategies. These real-world examples demonstrate the tangible impact of leveraging business intelligence tools in enhancing decision-making processes. However, this integration is not without its challenges. E-commerce businesses face hurdles such as data integration issues, ensuring data accuracy, and the need for skilled personnel to interpret and act upon insights generated. Navigating these challenges presents opportunities for innovation and refinement in the application of business intelligence for E-commerce churn modeling. By addressing these challenges and capitalizing on opportunities, businesses can harness the full potential of business intelligence tools, fostering a proactive and adaptive approach to churn management in the dynamic and competitive landscape of E-commerce.

#### **2.10.3. Methodology**

##### **Data Collection and Preprocessing**

The data collection and preprocessing phase focused on utilizing the "Ecommerce Customer Churn Analysis and Prediction" dataset created by Ankit Verma [4]. Governed by a CC BY-NC-SA 4.0 license, this dataset granted researchers permission to share, copy, and redistribute the material under the condition of providing appropriate credit, linking to the license, and specifying any modifications made. The original dataset was in `xlsx` format, which was converted to `csv` for better management.

To align the dataset with our research objectives, unnecessary columns were removed, and each column represented a different customer characteristic. These included `PreferredLoginDevice` (string), `PreferredPaymentMode` (string), `Gender` (string), `PreferredOrderCat` (string), `SatisfactionScore` (int 0-5), `Complain` (int 0-1), `CouponUsed` (int), `OrderCount` (int), `CashbackAmount` (float), and `Churn` (int 0-1). These variables encompassed both categorical and numerical features, providing a comprehensive view of customer behavior.

Following data selection, rigorous cleaning techniques were applied, involving column renaming, removal of empty rows, and elimination of special characters that could impact the accuracy of algorithmic predictions. This meticulous data preparation aimed to enhance the quality and reliability of the dataset for robust churn predictions.

The resulting dataset underwent random splitting into training (70%), validation (15%), and test (15%) sets. However, an issue arose during the random splitting process where many labels identified during prediction were not present in the training or validation sets. To address this, the `stratify` parameter in the `train_test_split` function was utilized, ensuring that the distribution of labels was similar in the training and validation/test sets. This approach aimed to maintain a representative sample for effective model training and evaluation.

## Logistic Regression and Neural Networks Training and Predictions

The main question of the research aimed to determine whether the Logistic Regression model is more efficient than the Neural Network after training for churn prediction in e-commerce. For the purposes of the study, the dataset was uploaded to Google Drive, and Google Colab/Jupyter with a CPU was utilized to train and run the models.

In the process of training a logistic regression model for churn prediction in the E-commerce domain, the datasets, comprising both training and validation sets, are seamlessly loaded from Google Drive using `Pandas`. The categorical variables undergo transformation into dummy variables via one-hot encoding, streamlining the model training process. Following this preprocessing step, a logistic regression model is instantiated and effectively trained on the specified training dataset. The model's convergence is ensured with a maximum of 1,000 iterations, and reproducibility is guaranteed by setting a specific random seed. Subsequently, the adeptly trained model is applied to make predictions on the validation set, and its performance is meticulously evaluated using the accuracy score metric. To ensure the preservation and accessibility of this valuable model, it is conscientiously saved in the designated Google Drive directory for potential future use.

In the pursuit of training a neural network for churn modeling within the E-commerce domain, the initial step involves mounting Google Drive for efficient file access and storage. Subsequently, the training and validation datasets are seamlessly loaded using `Pandas`. The features and target variable are then meticulously separated, and categorical variables undergo transformation into dummy/indicator variables through one-hot encoding. For enhanced model performance, numerical features are standardized. The construction of the neural network entails a sequential architecture with two hidden layers, comprising 64 and 32 nodes, respectively, and a final output layer employing a sigmoid activation function for binary classification. The model is compiled using binary crossentropy loss and the Adam optimizer. Training unfolds over 10 epochs with a batch size of 32, utilizing validation data for robust generalization assessment. Following training, predictions are generated on the validation set, and the model's accuracy is evaluated using `scikit-learn`'s `accuracy_score` metric. The adeptly trained neural network model is then saved as an `HDF5` file in the designated Google Drive directory, ensuring its accessibility for future use. Notably, the model's predictive capabilities extend to making predictions on the test set, thereby completing the comprehensive churn modeling process.

To facilitate transparency and reproducibility, all relevant resources, including the dataset, the data cleaning process, the split datasets, the complete code, and Jupyter Notebook (`ipynb`) file, have been made available on GitHub under an MIT license. This comprehensive sharing of resources aims to encourage collaboration, scrutiny, and further exploration within the research community [232].

### 2.10.4. Results and Discussion

#### Understanding E-commerce Customers

In our e-commerce dataset, we have identified that 19.79% of male customers and 14.69% of female customers are susceptible to churn. This initial exploration of gender-specific churn rates serves as a foundational step for in-depth analysis and the development of targeted retention strategies for both male and female customer segments.

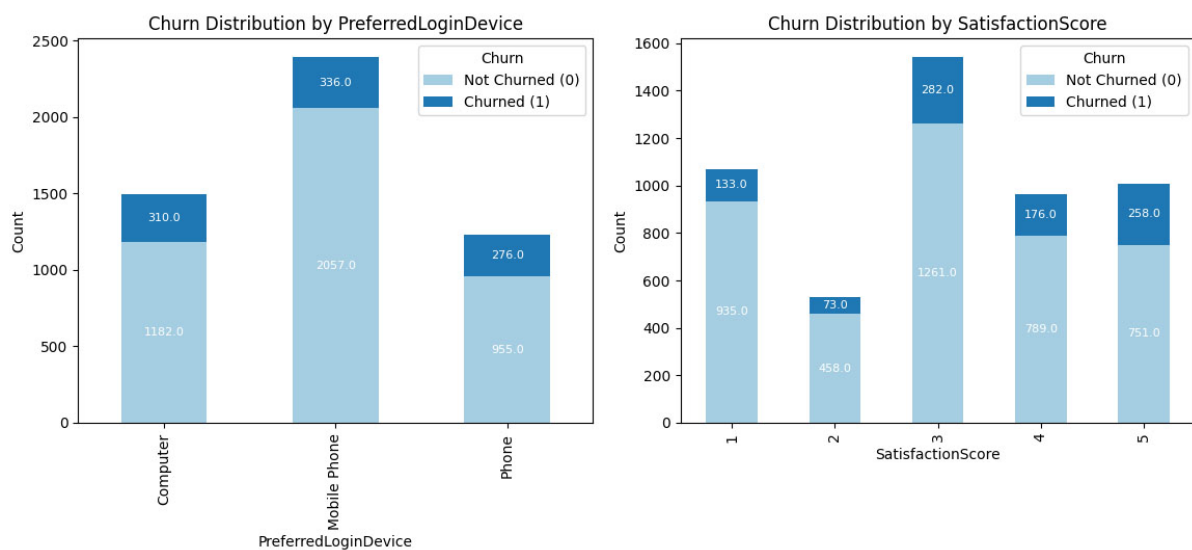
Examining churn rates in relation to payment methods, we found variations, such as 26.98% for credit card, 25.68% for cash on delivery, 8.33% for cash on delivery, 15.88% for credit cards, 14.96% for debit cards, 24.56% for e-wallets, and 17.35% for UPI. This breakdown sheds light on customer preferences and reveals potential areas for improvement within the payment processing system.

Moreover, the probability of churn is influenced by the ordering method, with 25.27% experiencing churn through phone orders, 18.86% through computer transactions, and 13.09% through mobile phone transactions. Recognizing the impact of ordering methods on churn rates underscores the necessity of optimizing mobile user experiences and tailoring strategies accordingly.

Delving into specific purchase categories, our analysis indicates that 30.23% of customers are likely to churn while purchasing mobile phones, 27.71% for general mobile purchases, 15.71% for fashion goods, 5% for grocery items, 8.9% for laptops and accessories, and 9.3% for other products. Understanding churn rates across different product categories empowers targeted marketing and retention efforts aligned with customer preferences.

Furthermore, our dataset unveils that 35.69% of customers who previously voiced complaints are prone to churn, while 10.58% of customers churn without registering any complaints. This highlights the critical importance of addressing customer concerns, as satisfied customers are less likely to churn. Establishing a connection between customer complaints and churn emphasizes the pivotal role of proactive customer service and resolution in maintaining overall customer satisfaction and loyalty.

Indicatively, in **Figure 16**, the results of the data analysis before training on the full dataset are presented for two of the customers' features/columns, namely, PreferredLoginDevice and SatisfactionScore.



**Figure 16.** Churn Distribution by PreferredLoginDevice and SatisfactionScore

### 2.10.5. Performance Comparison of Logistic Regression and Neural Networks

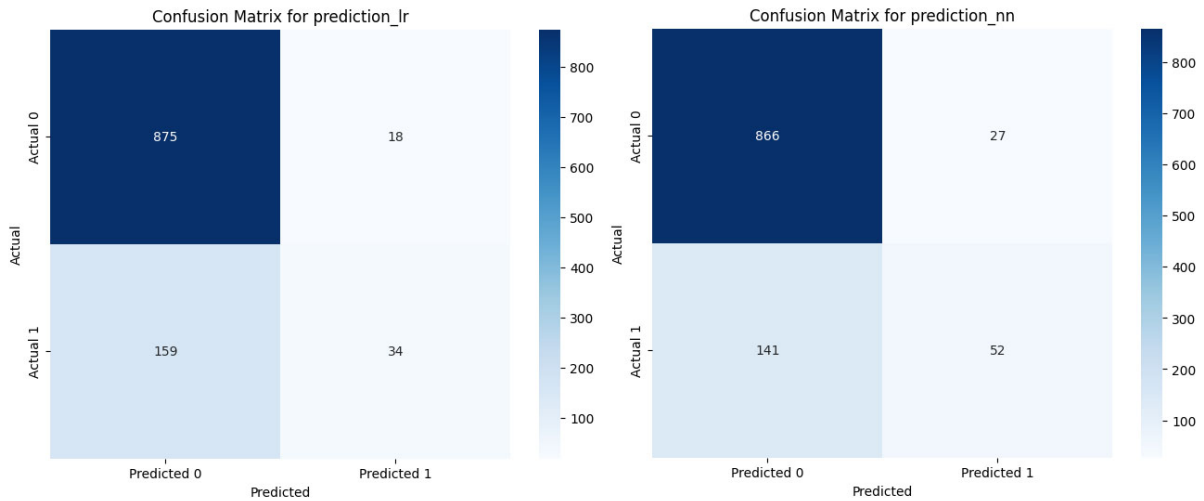
Following the training of logistic regression and neural network models with the training and validation sets, these trained models were utilized to predict customer churn based on features within the test set. The resulting predictions were stored in distinct columns, namely, prediction\_lr and prediction\_nn. To evaluate the precision of each model's predictions, we employed the confusion\_matrix and classification\_report methods from the sklearn library. In this analysis, each prediction was compared with the actual churn value, and the outcomes are presented in **Table 6** and **Table 7**, with the corresponding confusion matrices in **Figure 17**.

**Table 6.** Accuracy of the predictions made by the Logistic Regression-trained model.

	Precision	Recall	F1-Score	Support
0	0.85	0.98	0.91	893
1	0.65	0.18	0.28	193
accuracy			0.84	1086
macro avg	0.75	0.58	0.59	1086
weighted avg	0.81	0.84	0.80	1086

**Table 7.** Accuracy of the predictions made by the Neural Network-trained model.

	Precision	Recall	F1-Score	Support
0	0.86	0.97	0.91	893
1	0.66	0.27	0.38	193
accuracy			0.85	1086
macro avg	0.76	0.62	0.65	1086
weighted avg	0.82	0.85	0.82	1086



**Figure 17.** Confusion matrices for predictions made by the trained models

Upon reviewing the data in **Table 6** and **Table 7**, it is evident that the logistic regression trained model achieves an overall precision of 81%, while the neural network trained model attains a precision of 82%. The overall precision is calculated as the weighted average of precision values for each class, considering their respective support. This information is outlined in the weighted avg line of the classification reports.

The comparison indicates that the neural network trained model exhibits slightly higher precision in its predictions compared to its counterpart, the logistic regression trained model. Both models demonstrate high effectiveness in this task, especially considering the relatively small size of the training and validation set, which comprises approximately 4,092 rows.

However, it is crucial to highlight the limitations of the dataset, particularly in the training set, where only 533 out of 3,069 customers have churned. The scarcity of churned customers in the training set restricts the models from accurately predicting churn for these specific cases, underscoring the importance of a balanced dataset for achieving more accurate predictions post-training.

The underlying concept of churn modeling in e-commerce holds immense significance for e-commerce owners. Knowing in advance whether a customer is likely to churn, based on their unique characteristics and previous purchases, is crucial. By fine-tuning their e-commerce strategy using the insights derived from the trained models in this study, any e-commerce owner can predict and proactively mitigate churn.

This proactive approach to churn prediction not only allows e-commerce owners to retain valuable customers but also contributes to the overall success and sustainability of their business. Understanding the factors that contribute to customer churn provides a strategic advantage, enabling owners to tailor marketing efforts, enhance customer experiences, and address potential issues before they escalate.

Furthermore, the insights gained from the trained models empower e-commerce owners to personalize interactions with customers, offering targeted promotions or incentives to those identified

as at-risk for churn. This level of personalized engagement not only fosters customer loyalty but also helps optimize resources by directing marketing efforts towards segments with a higher likelihood of conversion.

#### **2.10.6. Section Conclusion**

In the dynamic e-commerce landscape, characterized by substantial customer acquisition costs, the strategic imperative of averting churn emerges as a cost-effective necessity. The capacity to foresee and address potential churners not only translates into heightened customer lifetime value but also contributes to an enhanced brand reputation and sustained profitability. This study employed two well-established models trained on a specialized dataset tailored for churn modeling. These models were tasked with making predictions to discern their respective precision levels. The results revealed both models to be highly accurate, with the neural network trained model slightly surpassing the logistic regression trained model in precision. Additionally, the study underscored the pivotal role of dataset quality and balance in the training process. Beyond the realm of predictions, this research furnishes e-commerce owners with more than just insights; it offers a proactive roadmap for customer relationship management, paving the way for enduring business success.

### 3. Web Accessibility and Progressive Web Apps (PWA)

*Adapted from:*

Roumeliotis, K.I.; Tselikas, N.D. *Evaluating Progressive Web App Accessibility for People with Disabilities. Network* 2022, 2, 350-369. <https://doi.org/10.3390/network2020022>

#### 3.1. Introduction

An increasing number of users are interacting with websites and mobile apps daily, and mobile apps are experiencing tremendous growth, with users appreciating the usability and the user experience (UX) they offer. The growing users' need for mobile apps has led to the discovery of hybrid apps that combine web technologies and native functions. However, both hybrid and native apps have some limitations, such as a free space commitment on devices and more difficult multiplatform upgrades. In 2015, Google provided a solution to the above-mentioned restrictions with the introduction of a new technology called Progressive Web Apps (PWAs). PWAs bring features we expect from native apps to the mobile browser [14]. They are built and enhanced with modern application programming interfaces (APIs) to deliver enhanced capabilities, reliability, and installability while reaching anyone, anywhere, and on any device with a single codebase [15].

At a time when web technologies are rapidly evolving, developers' interest in accessibility is fading. The need concerning people with disabilities for equal access to new technologies is growing. Even academic and government websites fail to offer real accessibility, thus making the lives of people with disabilities difficult [233].

Although much research focuses on the performance evaluation and Search Engine Optimization (SEO) of PWAs, their corresponding efficiency in terms of accessibility guidelines' conformance has not been investigated to date [12,14,15,234,235]. This study presents a review of the existing literature on PWA and accessibility technologies, and it is motivated by the need for accessibility solutions among platforms. The contributions of this research are summarized as follows:

- We include a discussion of PWAs and accessibility challenges as a guide for future research.
- We extensively review PWAs as a technology and articulate some drawbacks that arise.
- We critically investigate the accessibility technologies to date, including the newly introduced as a working draft Web Content Accessibility Guidelines (WCAG) 2.2.
- We combine and evaluate PWAs and accessibility technologies to promote web accessibility for anyone.

#### 3.2. Materials and Methods

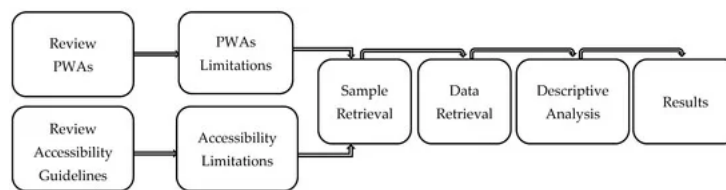
The purpose of this study is to evaluate the PWAs' effectiveness against the Accessibility Guidelines and to promote the necessity of a more accessible web.

We followed six research stages to achieve the desired outcomes:

- i. Literature review. Both PWAs and Accessibility Guidelines have been reviewed summarizing the existing literature.
- ii. Critical review. Both PWAs and Accessibility Guidelines address certain limitations that complicate the overall user experience (UX). From a critical point of view, we summarize those limitations.
- iii. Sample retrieval. In this stage, a sample of 20 PWA and 20 non-PWA websites are discovered and placed on a table.
- iv. Data retrieval. Each of the websites was fully tested against 10 accessibility evaluation tools—auditing tools. The results from the accessibility evaluation were added to a table.

- v. Descriptive analysis. In this stage, using the Jupyter Notebook software, we made the descriptive analysis of our dataset.
- vi. Presenting the results of the descriptive analysis.
  - a. Evaluation tools' limitations. After the descriptive analysis, the assumption that each accessibility evaluation tool performs its own measurements was confirmed.
  - b. PWAs' accessibility limitations. It was also confirmed that PWAs are not fully accessible by default and that web developers must make an effort to conduct both machine and manual audits to achieve actual accessibility.
  - c. PWAs' overall performance. In this stage, we conclude that PWAs are more accessible than conventional websites, observing that they have a lead in performance and SEO.

**Figure 18** summarizes the methodology that has been used to conclude valuable outcomes regarding the PWAs' accessibility performance.



**Figure 18.** Overall step-by-step representation of the proposed methodology, reflecting the stages for evaluating PWAs' accessibility.

### 3.2.1. Research Limitations

Since PWAs are quite new as a technology, they also have a low adoption rate by web developers. To avoid inaccurate results from websites that have partially or wrongly adopted PWAs, for our analysis, we used the PWA websites presented as case studies on the official Google page for PWAs. For each PWA website of our dataset, we chose a regular non-PWA website of the corresponding niche and industry (top-listed in Google search results).

The online tools used for the purposes of this research are:

- Lighthouse by Google;
- Wave by WebAim;
- Web Accessibility by Level Access;
- Power Mapper;
- CSS Validation Service by W3;
- Nu Html Checker by W3;
- Axe DevTools Accessibility;
- Insights Chrome;
- Equal Access by IBM;
- Site Improve.

### 3.3. Progressive Web Apps (PWA)

Prior research has found that almost four fifths (79%) of internet users access the internet on their mobile devices [236]. Statistical research agencies support the statistics that mobile internet traffic as a share of total global online traffic is greater than 55.64% and that, in 2020, mobile app downloads worldwide amounted to 218 billion [237].

Mobile devices are an indispensable part of human life nowadays. Mobile apps have been developed for every field and any need, from taking and editing photos to accessing social media, attending online meetings, and more. The incremental need for mobile apps led development companies to develop native or hybrid applications to overcome the limitations that the web as a

platform imposes on mobile devices [14]. However, even mobile apps face some limitations, such as the following:

- Platform-specific applications;
- Devices' resources consumption;
- Multiplatform updates;
- SEO;
- Challenging app store optimization.

To overcome these limitations, in 2015, Google introduced a new technology, namely, PWAs, which are web applications built with modern APIs to deliver enhanced capabilities, reliability, and installability while reaching anyone, anywhere, and on any device with a single codebase [15]. As browsers become more modern, an increasing number of features become available to end users. This is known as progressive enhancement [12]. PWAs try to solve UX and user offline use. Moreover, they combine the advantages of apps and the web, such as being progressive, reliable, responsive, offline functional, interactive, installable, and notifiable [237]. PWAs are based on three main pillars:

- Capable: new and upcoming APIs are more capable than ever, introducing new capabilities, from file system access to app badging.
- Reliable: PWAs are reliable, fast, and dependable, regardless of network speed.
- Installable: PWAs run as apps on a user's home screen without any browser tab.

Furthermore, PWAs' main characteristics are as follows:

- Progressive: using service workers and a web app manifest, PWAs work on every modern browser.
- App-like: they use the app-shell model to provide app-style navigations and interactions [13].
- Installable: installed PWAs appear on the launch surfaces of any devices, such as the Mac OS X Applications folder and Spotlight function.
- Independently installed: they bypass App Store and Google Play installations.
- Responsive: PWAs are responsive and accessible from any platform and browser.
- Independent of connectivity: they are reliable even under unstable network conditions.
- Capable of delivering an offline experience: using service workers and the IndexedDB storage system, PWAs can provide offline experience.
- Safe: PWAs are served over Transport Layer Security (TLS) to ensure a secure connection.
- App badging: PWAs allow badging for app icons to subtly notify users of a new activity that might require their attention [15].
- Capable of supporting custom offline pages and splash screens: like native apps, PWAs support offline pages and splash screens.
- Fresh: PWAs are always kept up to date using the service worker update process.
- Discoverable: PWAs following W3C manifests are more likely to be discovered by search engines.

### 3.3.1. Service Workers

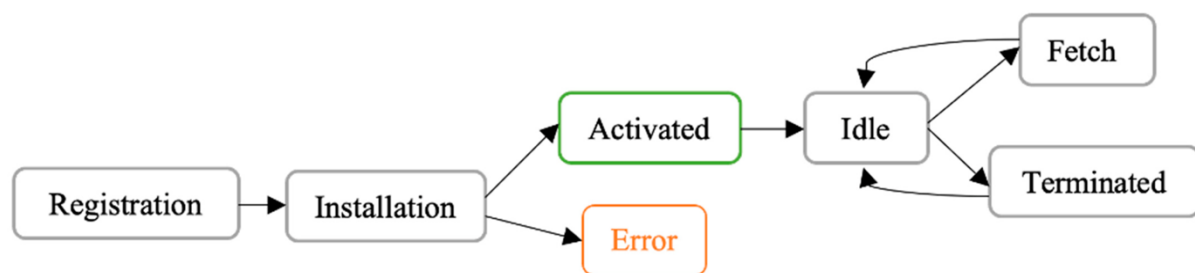
A service worker is a set of APIs introduced by PWA, running in its own thread and providing generic entry points for event-driven background processing that allows developers to programmatically cache and preload assets and data and manage push notifications, among other things [238]. More specifically, a service worker is a JavaScript script that runs in the background of the application, deploying in a separate thread from the UI, thereby avoiding app freezing. It acts as an intermediary between the app and the internet [234], and its main purpose is to execute functions as "promises," in a specific order, sending results back to the app. A JS Promise is an asynchronous managing mechanism that enables programmers to chain asynchronous computations while supporting proper error-handling methods [239]. Promises technically fix the gap between function execution order by telling the asynchronous method that it "promises" to call a given function as soon

as the asynchronous (async) function is finished. An alternative method to promises is the callbacks. As regards the callbacks, the executing function knows in advance what has to be done when the asynchronous task has been completed. In promises, instead, the executing function returns an object (the promise), in which we describe what it has to do when the asynchronous task has been completed. A general approach is that the callbacks follow a more functional way of programming, instead of promises following a more object-oriented way.

Before service workers, the legacy technology called AppCache was the one that provided users an offline experience on the web. Although AppCache API was a straightforward solution to the offline experience, it faces a lot of issues that service workers avoid, such as caching all pages by default whether they have changed or not.

A service worker, as a middle service, is vulnerable to cross-site scripting (XSS) attacks, and a secure connection is, thus, required [240]. In addition, service workers have the power to provide async capabilities to an app, such as periodic sync, and act similarly to a cronjob performing specific actions at a certain date or time.

To set up a functional service worker, three main actions must be implemented: registration, installation, and activation. Through an on-page load, service workers must check for browser compatibility. Then, using the `serviceWorkerRegistration` function, the installation starts as a promise [241], where the `install` and `activate` `AddEventListener` functions follow a specific procedure to install and activate the service workers. The process followed by developers to set up a service worker is presented in **Figure 19**.



**Figure 19.** Service worker's lifecycle.

Modern web browsers have adopted the use of service workers, and others support them in their latest versions. Supported web browsers are Chrome, Firefox, Opera, Safari, IE, and Samsung Internet. These browsers also support features such as promises; `navigator.serviceWorker`; and `fetch`, `install`, and `activate` events. In contrast, background sync is currently only supported in Chrome [242]. According to Google's Web Updates in 2015, background sync is a web API that lets you defer actions until the user has stable connectivity. This is useful for ensuring that whatever the user wants to send is actually sent.

### 3.3.2. Caching Storage Application Programming Interface (API) and Offline Functionality

The service worker kernel is equipped with a caching interface; the cache can store static content that users can access without an internet connection. Depending on the caching settings that have been set, the service worker undertakes caching of the content that the user has already visited. If the user's device does not have an internet connection, all cached information is temporarily displayed until the connection is restored [243]. In that way, PWAs achieve the fastest loading of apps, making the apps usable even without an internet connection. The only limitation is the browser's storage limit. The process of creating and retrieving caching content is presented in **Figure 20**.

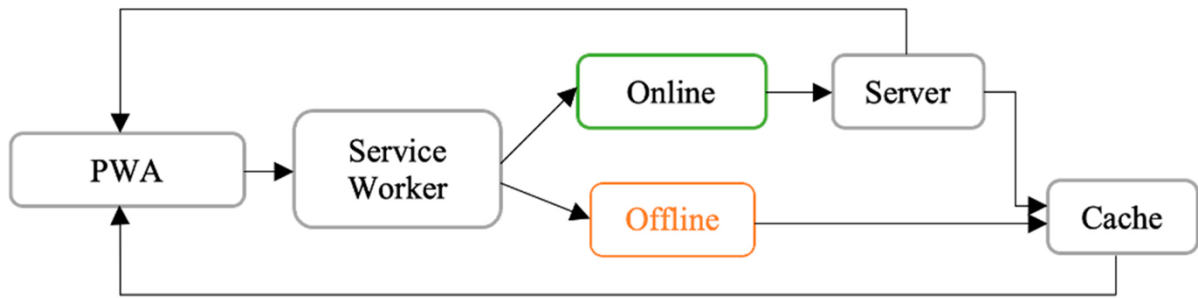


Figure 20. Caching mechanism

To set up a caching mechanism, an `addEventListener` must be installed. Through the installation, event items are added on the cache's property as an array, and these properties can be used instead of the server's content when there is no internet connection. Moreover, even if an internet connection exists, the service worker sends the cached content back to the app, and only the missing content is requested by the webserver. To keep the cached content up to date, the service worker must regularly update its cached contents by deleting old caches and creating a new one. If the page is dynamic and its content changes frequently, refreshing the dynamic cached elements is possible with periodic JSON fetch calls.

Adopting techniques from native apps, PWAs suggest creating an offline fallback page with a custom service worker. On this page, the user is notified of the lack of internet connection and, with an automatic reload based on the online event, is informed that a new reconnection attempt will be performed in x seconds [15].

### 3.3.3. Web App Manifest

The web app manifest is a JSON file that informs the browser about the PWA and how it should behave when installed on the user's desktop or mobile device [15]. This file contains the PWA's properties in JSON format. The properties required for a web app to be considered as a PWA are presented in Table 8, although many other optional properties exist that define more accurately the web app manifest [15].

Table 8. Web manifest properties.

Property	Description
<b>short_name and name</b>	These properties define the app's name shown under the app icon.
<b>icons</b>	These properties present the app icons in different sizes. For each icon, src, type, and sizes must be included.
<b>start_url</b>	This property describes the browser, which is the starting URL of the app.
<b>background_color</b>	This property sets the background color of the app.
<b>display</b>	This property is the display option. There are four display options: full-screen, standalone, minimal-ui, and browser. Each of these options displays the app differently.
<b>scope</b>	This property is a string that specifies which directories and files the web app manifest affects.
<b>theme_color</b>	This property defines the color of the toolbar.
<b>shortcuts</b>	This property is an array of app shortcut objects whose goal is to provide quick access to key tasks within the app. For each icon, name and URL values must be included.
<b>description</b>	This property is a short description of the app's scope.
<b>screenshots</b>	This property is an array of image objects, representing the app as a screenshot in common usage scenarios. For each image, src, sizes, and type must be defined.

Once the JSON file is created, it should be added as a link to the head of each page. A file validation using Chrome's dev tools is also deemed necessary.

Acting as native apps, PWAs can be added to users' home screens so that the users do not have to navigate to a URL every time they want to use an app [244]. Furthermore, through the launching process, instead of a blank white screen, a splash screen is displayed [12]. Chrome for Android automatically shows a custom splash screen as long as the PWA meets some basic requirements, such as name, background color, and icons, in its web app manifest [15].

### 3.3.4. Push API and Notifications

PWAs have the ability to display re-engaging notifications as defined in the Push API [244]. On desktops, Chrome, Firefox, and Opera all support both the Push API and web notifications. While Safari supports web notifications, it has a custom implementation for push notifications too. Moreover, Edge supports web notifications, but has no Push API support. On the mobile side, iOS has no support for either feature, while Android supports only the Push API [12].

### 3.3.5. Application Shell (App Shell) Architecture

An application shell (app shell) refers to the local resources required by a web app to load the skeleton of the user interface (UI) [235]. After native app implementation, all of the views, fonts, and images are uploaded and rendered as a basic app skeleton on the app store. In contrast, PWAs are not uploaded to the native app store but are fetched at runtime when the app is opened [245]. The app shell contains HTML, JavaScript, and CSS files that change infrequently and can be cached so that they can be loaded instantly from the cache on repeat visits. If the app shell has been cached by the service worker, then, on repeat visits, the app shell allows users to rapidly receive meaningful pixels on their screens without a network. UI pieces across different pages, such as headers, toolbars, and footers, are commonly fetched and cached on the app shell. The main purpose of the app shell is to make a PWA's users feel as though they are using a "real" app [15].

The app shell should:

- Load fast;
- Use as little data as possible;
- Use static assets from a local cache;
- Separate content from navigation;
- Retrieve and display page-specific content (HTML, JSON, etc.).

### 3.3.6. PWA Critical View

This study identifies and evaluates, through testing, all of the advantages and disadvantages of PWAs that have emerged to date. These advantages and drawbacks are identified and thoroughly analyzed in **Table 9**.

Based on **Table 9**, we conclude that PWAs are fully compatible only with Android devices, thus depriving iOS users of many default features. However, PWAs are not responsible for the majority of these limitations; Apple's operating systems have many safety valves to enhance the security of their users, but set limits on innovation. Furthermore, PWAs constitute a new, rapid technology, and dealing with the limitations that arise takes time. Updates are frequent, making PWAs closer to native functions. For this technology to evolve, both developers and operating system must work together with browser companies.

**Table 9.** Advantages and disadvantages of PWAs.

Advantages	Disadvantages
PWAs are cross-platform solutions. They are created once for all platforms.	PWAs cannot be deployed to the App Store or Google Play. Nonetheless, third-party technologies can convert PWAs into native apps using a protocol based on Custom Tabs.

PWAs allow for quick installations without waiting times.	PWAs are not fully supported by iOS and OsX devices. Many features, such as push notifications for iOS, cannot be achieved without a developer registering an app in the Apple developer portal.
The cost to develop a PWA is much lower than a native or hybrid app.	The add-to-home-screen feature is not available for iOS. Users must manually add the app to the home screen through settings options.
The cost to maintain a PWA is much lower than a native or hybrid app.	PWAs cannot access native iOS components, such as Face ID and Bluetooth.
PWAs are installable. Installing a PWA allows it to look, feel, and behave like all other installed apps.	PWAs are not allowed to access Apple's iBeacons, depriving them of using phone battery and altimeter features.
PWAs can be updated on the fly without a user's interaction.	PWAs have no access to the iOS local filesystem, contact book, and current location.
PWAs provide a custom offline page.	Background sync, which is a core web API, is only supported by the Chrome browser.
PWAs are reliable, fast, and dependable regardless of the network speed.	Web developers must calculate the caching limits for each browser to offer a decent offline caching experience.
Chromium-based PWAs can access the hardware features on Android devices in the same way as native mobile applications.	Face and Speech Recognition is only available through third-party APIs.
PWAs are equally usable with a mouse, a keyboard, a stylus, or touch [15].	Only in-app purchases can be used, not payments for each app download.
	PWAs are incompatible with old devices' obsolete browsers.
Well-known platforms, such as WordPress and Magneto, have already implemented plugins and templates to easily develop PWAs.	PWAs escape the app approval process, and low-quality web apps will eventually emerge.
A PWA can be used as a light version of an app for devices with fewer resources.	Ratings, reviews, and responses on app stores are indispensable in helping users to select the most appropriate app depending on their needs. Unfortunately, PWAs are not a part of app stores.
PWAs can be accessible by meeting WCAG standards.	The mobile-first approach that PWAs often promote will end with mobile-only apps. PWAs' templates are not as desktop-friendly.
PWAs can interact flawlessly with powerful APIs.	Using more internet resources compared with native apps, PWAs consume more battery.
A PWA can be properly indexed on search engines, promoting itself using search engine optimization techniques.	As web applications, PWAs are more likely to be hacked.
PWAs can expand the number of returning visitors, increase conversion rates and engagement, and reduce data usage. Google's case studies have shown that AliExpress increases the conversion rate of new users by 104%, with new PWAs and the Twitter Lite PWA significantly increasing pages per session by 65%, and exhibiting a 75% increase in Tweets sent and a 20% decrease in bounce rate [246].	PWA developers use pop-ups to alert users that can add the app to their home screen. However, some pop-up notifications may be blocked by modern browsers.

### **3.4. Web Accessibility**

Web accessibility allows everyone, including people with disabilities, to perceive, understand, navigate, and interact with the internet [247]. The same applies to mobile apps and PWAs. Web accessibility can be defined as making a website navigable and tractable through various user categories, especially users who have disabilities and normally face obstacles and limitations when interacting with the web via electronic devices [248]. Web accessibility is related to the practice of generating web pages accessed by people with all types of abilities and disabilities [249]. Despite technological growth and international regulations, the majority of websites and apps remain inaccessible to certain groups of people. Although accessibility seems to be obsolete, an increasing number of new technologies are relying on its guidelines to create user-accessible applications, such as voice search and PWAs.

#### **3.4.1. Guidelines and Regulations**

The growing number of national and international laws addressing the accessibility of information and communication technologies (ICT), including the web, has resulted in many different approaches in practice [248]. The first accessibility law adopted by the US in 1990 is the Americans with Disabilities Act (ADA). According to the law, websites and apps are deemed to be places of accommodation and have the duty to be accessible; those that are not accessible are considered to discriminate against people with disabilities [250].

There is a wide range of people with disabilities, including visual, auditory, physical, speech, cognitive, language, learning, and neurological disabilities, and combinations thereof [251]. While numerous guidelines and tools have been developed to help improve access to and understanding of websites' content, the most relevant ones are W3C WCAG, ISO, and Section 508 [252].

Although this study focuses on W3C WCAG guidelines, it is imperative to mention shortly ISO 9241-151:2008 and Section 508 regulations. With the aim of increasing usability, ISO 9241-151:2008 – Guidance on World Wide Web UIs provides guidance on the human-centered design of software web UIs [253]. ISO 9241-151 focuses on the design aspects and provides design guidance and recommendations in four major areas, which are [252,254]:

- High-level design decisions and design strategy;
- Content and functionality;
- Navigation, search, and interaction;
- Media design and presentation.

In 1998, US Congress amended the Rehabilitation Act of 1973 to require federal agencies to make their electronic and information technology (EIT) accessible to people with disabilities [255]. Section 508 of the Rehabilitation Act applies to the federal agencies, and it requires e-government websites to be accessible to people with disabilities [252]. Section 508 complies with the requirements of other guidelines and standards for both the US and the European Commission, as well as the W3C WCAG.

#### **W3C Web Content Accessibility Guidelines (WCAG)**

The World Wide Web Consortium (W3C) has developed WCAG to make the web accessible to people with disabilities [256]. Following these guidelines will make content more accessible to a wider range of people with disabilities. Making the web accessible benefits individuals, businesses, and society [251], and the WCAG's goal is to provide a single shared standard for web content accessibility that meets the needs of individuals, organizations, and governments internationally [256].

Four different WCAG versions exist: WCAG 1.0, WCAG 2.0, WCAG 2.1, and WCAG 2.2. Each version extends its predecessor and takes its requirements for granted.

WCAG 1.0, created by W3C on 5 May 1999, is the foundation stone of WCAG regulations. This first version consists of 14 guidelines or general principles of accessible design. Each guideline has a unique statement, a guideline number, and checkpoints. Respectively, the checkpoints have a unique

statement, a checkpoint number, and a link to a section of the Techniques Document where implementations and examples of the checkpoint are discussed. To categorize the checkpoints depending on their importance each checkpoint has a priority level assigned by the Working Group based on the checkpoint's impact on accessibility [257]. In this process, there arose three levels of conformance:

- Conformance Level "A": all Priority 1 checkpoints are satisfied;
- Conformance Level "Double-A": all Priority 1 and 2 checkpoints are satisfied;
- Conformance Level "Triple-A": all Priority 1, 2, and 3 checkpoints are satisfied.

Although the conformance level was created to identify the priority of each checkpoint, at the same time, as the level increases, it is difficult for web developers to achieve the corresponding requirements, meaning that A-level checkpoints are easier to achieve than the Triple-A.

WCAG 2.0 came as an update to WCAG 1.0 in 2008. They improved the structure and consistency of the prior guidelines and, at the same time, added additional success criteria. The checkpoints were renamed to success criteria and the priority was assimilated from Level A to AAA. WCAG 2.0 is organized around four design principles that provide the foundation for web accessibility (perceivable, operable, understandable, and robust) [248].

- Level A is the minimum level that defines the foundations of web accessibility [258]. It consists of 12 guidelines, including 25 success criteria.
- Level AA includes all Level A and AA requirements [251]. In addition, it attaches 13 new success criteria to the existing six guidelines. These success criteria provide the basic objectives that authors must achieve to create more accessible content for users with different levels of disability [30].
- Level AAA includes all Level A, AA, and AAA requirements [251]. Moreover, it attaches 23 new success criteria to the existing nine guidelines. These success criteria allow for the evaluation of requirements and needs, such as design specifications, purchasing, regulation, or contractual agreements [258].

WCAG's 2.1 updated version came in June 2018. WCAG 2.1 came as an extension of the 2.0 version, adding five A, seven AA, and five AAA success criteria to the existing guidelines.

The newest WCAG 2.2 was introduced as a working draft in August 2020 and updated in May 2021. WCAG 2.2 extends its predecessor's guidelines, adding the following nine success criteria [259]:

- Accessible Authentication (3.3.7; Level A).

If an authentication process relies on a cognitive function test, then at least one other method must also be available that does not rely on a cognitive function test [259]. Cognitive deficits usually occur in older people [260], who find it difficult to memorize passwords, make calculations, or even solve a puzzle. Since this success criterion is required for conformance to Level A, there must be an alternative method in case the website uses a cognitive authentication function.

- Redundant Entry (3.3.8; Level A).

Redundant entry is a success criterion of Level A, which requires auto-populated or user-selected information in fields previously entered by the user. However, there is an exception when re-entering essential information [259].

- Consistent Help (3.2.6; Level A).

When facing problems completing a task on a website, people with some types of disabilities may not be able to work through the issue without further help. Issues could include difficulty completing a form or finding a document or page that provides the information required to complete a task [259]. Therefore, up-to-date Frequently Asked Questions and human contact details, or a messaging mechanism is required.

- Page Break Navigation (2.4.13; Level A).

The purpose of this success criterion is to allow people using assistive technology or screen readers to find references to content based on the page break locators found in the default view or printed version of a publication [259].

- Dragging Movements (2.5.7; Level AA).

All functionality that uses a dragging movement for operation can be operated by a single pointer without dragging, unless dragging is essential [259]. Dragging movements are observed in services such as Google Maps on mobile phones, where more pointers are needed to drag a map.

- Target Size (2.5.8; Minimum; Level AA).

This success criterion belongs to the AA conformance level and aims to help users with hand tremors and those who have difficulty with fine motor movement to activate interactive areas, such as in-line links and pop-over content accurately. Visible Controls (3.2.7; Level AA).

The controls needed to progress or complete a process are visible at the time they are needed without requiring pointer hover or keyboard focus, or a mechanism is available to make them persistently visible [259]. Tasks such as on-mouseover user interactions can make it challenging for people with impaired memory and other cognitive and learning disabilities.

- Focus Appearance (2.4.11; Minimum; Level AA).

The purpose of this success criterion is to ensure that a keyboard focus indicator is clearly visible and discernible following the appropriate minimum area and contrast [259].

- Focus Appearance (2.4.12; Enhanced; Level AAA).

This success criterion is an extension of the previous one, extending the minimum area, increasing contrast, and excluding obscured elements.

All the updated WCAG’s design principles, guidelines, and success criteria are presented in depth in **Table 10**.

**Table 10.** WCAG design principles, guidelines, and success criteria.

1. Perceivable	2. Operable	3. Understandable	4. Robust
<b>1.1 Text Alternatives</b>	<b>2.1 Keyboard Accessible</b>	<b>3.1 Readable</b>	<b>4.1 Compatible</b>
(A) 1.1.1 Non-text Content	(A) 2.1.1 Keyboard	(A) 3.1.1 Language of Page	(A) 4.1.1 Parsing
<b>1.2 Time-based Media</b>	(A) 2.1.2 No Keyboard Trap	(AA) 3.1.2 Language of Parts	(A) 4.1.2 Name, Role, Value
(A) 1.2.1 Audio-only and Video-only (Prerecorded)	(AAA) 2.1.3 Keyboard (No Exception)	(AAA) 3.1.3 Unusual Words	(AA) 4.1.3 Status Messages
(A) 1.2.2 Captions (Prerecorded)	(A) 2.1.4 Character Key Shortcuts	(AAA) 3.1.4 Abbreviations	
(A) 1.2.3 Audio Description or Alternative (Prerecorded)	<b>2.2 Enough Time</b>	(AAA) 3.1.5 Reading Level	
(AA) 1.2.4 Captions (Live)	(A) 2.2.1 Timing Adjustable	(AAA) 3.1.6 Pronunciation	
(AA) 1.2.5 Audio Description (Prerecorded)	(A) 2.2.2 Pause, Stop, Hide	<b>3.2 Predictable</b>	
(AAA) 1.2.6 Sign Language (Prerecorded)	(AAA) 2.2.3 No Timing	(A) 3.2.1 On Focus	
(AAA) 1.2.7 Extended Audio Description (Prerecorded)	(AAA) 2.2.4 Interruptions	(A) 3.2.2 On Input	
(AAA) 1.2.8 Media Alternative (Prerecorded)	(AAA) 2.2.5 Re-authenticating	(AA) 3.2.3 Consistent Navigation	
(AAA) 1.2.9 Audio-only (Live)	(AAA) 2.2.6 Timeouts	(AA) 3.2.4 Consistent Identification	
<b>1.3 Adaptable</b>	<b>2.3 Seizures and Physical Reactions</b>	(AAA) 3.2.5 Change on Request	

(A) 1.3.1 Info and Relationships	(A) 2.3.1 Three Flashes or Below Threshold	(A) 3.2.6 Consistent Help [NEW]
(A) 1.3.2 Meaningful Sequence	(AAA) 2.3.2 Three Flashes	(AA) 3.2.7 Visible Controls (NEW)
(A) 1.3.3 Sensory Characteristics	(AAA) 2.3.3 Animation from Interactions	<b>3.3 Input Assistance</b>
(AA) 1.3.4 Orientation	<b>2.4 Navigable</b>	(A) 3.3.1 Error Identification
(AA) 1.3.5 Identify Input Purpose	(A) 2.4.1 Bypass Blocks	(A) 3.3.2 Labels or Instructions
(AAA) 1.3.6 Identify Purpose	(A) 2.4.2 Page Titled	(AA) 3.3.3 Error Suggestion
<b>1.4 Distinguishable</b>	(A) 2.4.3 Focus Order	(AA) 3.3.4 Error Prevention (Legal, Financial, Data)
(A) 1.4.1 Use of Color	(A) 2.4.4 Link Purpose (In Context)	(AAA) 3.3.5 Help
(A) 1.4.2 Audio Control	(AA) 2.4.5 Multiple Ways	(AAA) 3.3.6 Error Prevention (All)
(AA) 1.4.3 Contrast (Minimum)	(AA) 2.4.6 Headings and Labels	(A) 3.3.7 Accessible Authentication (NEW)
(AA) 1.4.4 Resize text	(A) 2.4.7 Focus Visible	(A) 3.3.8 Redundant entry (NEW)
(AA) 1.4.5 Images of Text	(AAA) 2.4.8 Location	
(AAA) 1.4.6 Contrast (Enhanced)	(AAA) 2.4.9 Link Purpose (Link Only)	
(AAA) 1.4.7 Low or No Background Audio	(AAA) 2.4.10 Section Headings	
(AAA) 1.4.8 Visual Presentation	(AA) 2.4.11 Focus Appearance (Minimum) [NEW]	
(AAA) 1.4.9 Images of Text (No Exception)	(AAA) 2.4.12 Focus Appearance (Enhanced) (NEW)	
(AA) 1.4.10 Reflow	(A) 2.4.13 Page Break Navigation (NEW)	
(AA) 1.4.11 Non-text Contrast	<b>2.5 Input Modalities</b>	
(AA) 1.4.12 Text Spacing	(A) 2.5.1 Pointer Gestures	
(AA) 1.4.13 Content on Hover or Focus	(A) 2.5.2 Pointer Cancellation	
	(A) 2.5.3 Label in Name	
	(A) 2.5.4 Motion Actuation	
	(AAA) 2.5.5 Target Size (Enhanced)	
	(AAA) 2.5.6 Concurrent Input Mechanisms	
	(AA) 2.5.7 Dragging Movements (NEW)	
	(AA) 2.5.8 Target Size (Minimum) (NEW)	

### 3.4.2. Web Accessibility Critical View

Prior research has queried diverse approaches to address accessibility guidelines. This research not only identifies those guidelines, but also presents the problems that web developers face when trying to follow them, as well as the limitations of online tools that check for accessibility.

The success of web-based applications depends on how well they are perceived by the end users [256]. Making websites and apps accessible is mandatory. Thus, for example, in every new technology that Google presents, it emphasizes the need to create human-centered web applications. Furthermore, web developers, motivated by the best search engine rankings, follow some of the accessibility guidelines. However, the majority of web applications that are created are not fully accessible, thereby creating accessibility issues for people with disabilities. While some laws oblige website owners in both the public and private sectors to create websites that are accessible to everyone, these laws do not apply unless a person with disabilities complains about noncompliance.

### **Limitations of Accessibility Guidelines**

Previous research has shown that, even though accessibility guidelines lay the foundations for a more accessible web, they also have some limitations. Evidence suggests that compliance with accessibility standards does not always guarantee a satisfying UX on the web [261]. Furthermore, guidelines are difficult to evaluate, even if a combination of human and machine audits are used [262]. In a study with intellectual disabilities participants, it was observed that, even if a website follows the W3C WCAG, users' satisfaction is not 100% [263].

### **Web Developers' Limitations**

From a developer's point of view, dealing with accessibility is a demanding task that requires much time following each of the guidelines and success criteria. To implement real accessibility, every web development agency should hire a special team that will test the accessibility of the application, both with online tools and manually. A lack of awareness, education, and motivation leads developers to create inaccessible websites.

### **Limitations of Machine Auditing and Plugins**

Online tools (machine audits) and plugins have been developed to help website owners and web developers to create more accessible websites. Using online evaluation tools, such as WAVE (by WebAim), the W3 validator (by the W3C), and Lighthouse (by Google), every website can be checked against the WCAG and success criteria. These tools scan a website's source code and highlight the corrections that must be made to make the website more accessible. Some of them use visual composers to highlight areas where errors occur, while others refer to the appropriate success criterion to understand exactly what needs to be corrected. Although machine evaluation tools offer a useful pathway, each of them presents different website errors compared to the others.

In addition, some plugins for popular open-source platforms promise full accessibility with a single click. These plugins make some changes to a website's source code without necessarily covering any guidelines or success criteria.

Based on the information presented in this section, we conclude that the need for accessibility is increasing; however, an increasing number of websites—even government websites, online libraries, and university websites—do not follow the guidelines. On the one hand, even though accessibility guidelines cannot cover every disability issue, they are improved day by day. On the other hand, web developers have a moral obligation to provide a website that is as accessible as possible. To achieve this, both manual and machine auditing must be performed.

## **3.5. Results**

In **Section 3.3**, PWAs were extensively reviewed, and we conclude that the adoption of this new technology is accompanied by both positive and negative aspects. Then, in **Section 3.4**, accessibility guidelines that have been created to date were analyzed in depth, along with the limitations they face. In this section, we combine and evaluate the efficiency of PWAs in relation to accessibility.

Google’s website states that PWAs are fully accessible and, on the same page, web developers are encouraged to perform manual and machine auditing to achieve accessibility [264]. The term “fully accessible” is used by a growing number of technologies to indicate that the core of that technology is accessible. Accessibility, however, cannot be applied to the core, but to areas to which people with disabilities have access (i.e., the front end). As a result, PWAs could be described as a technology that promotes accessibility.

The ultimate goal of this research is to compare and evaluate the PWA as a technology against other (non-PWA) web technologies regarding their compatibility with the existing accessibility guidelines in order to conclude whether the former is more accessible than the latter or not. In this stage, a sample of 20 PWA and 20 non-PWA websites and e-commerces are selected as depicted in the first column of **Table 11** and **Table 12**, respectively. Our dataset consists of two groups of websites. The first group (first column in **Table 11**) contains 20 websites that were built based on the PWA architecture. These websites are also listed as case studies by Google [246]. The second group (first column **Table 12**) contains 20 conventional non-PWA websites. To achieve uniformity in our results, for each PWA website, a non-PWA website was found in the same niche. The non-PWA websites examined in this study are top-listed in Google’s search results based on their niche.

**Table 11.** Accessibility compliance errors for PWA websites

Dataset/Auditing Tools	T1 (%)	T2	T3	T4 (%)	T5	T6	T7	T8	T9	T10	Niche
Nikkei.com	18	102	88	15	5	1000	450	56	55	115	News Blog
George.com	11	2	10	17	30	63	49	10	18	62	Clothing Brand
Ele.me	12	1	2	19	0	5	23	7	10	30	Food Ordering
BookMyShow.com	4	6	127	26	11	851	252	233	242	32	Ticketing
Forbes.com	11	6	38	26	12	69	33	12	7	30	News Blog
Infobae.com	3	1	6	17	11	21	123	7	257	214	Digital-only news
Lancome-usa.com	8	1	13	14	72	154	166	0	85	190	Luxury cosmetics
Makemytrip.com	14	4	1	23	5	7	161	22	22	199	Travel Booking
Mynet.com	8	3	8	17	0	54	353	23	290	257	News Blog
Olacabs.com	22	166	1	33	1	50	150	83	163	139	Car Rental
Olx.in	33	13	1	30	5	65	50	40	78	52	Free Classifieds
Twitter Lite	26	11	2	0	1	24	53	38	34	40	Social Network
Wego.com	11	2	19	12	0	3	386	90	162	190	Travel Booking
Housing.com	41	12	9	12	4	26	117	26	22	54	Property Website
Alibaba.com	51	153	171	4	23	2	286	185	270	371	Marketplace
Weather.com	12	2	6	26	2	129	21	4	9	162	Weather Platform
Carnival.com	25	9	19	35	54	185	130	28	39	80	Travel Booking
Washingtonpost.com	13	39	5	0	17	254	65	47	58	64	News Blog
Aliexpress.com	58	26	46	2	0	16	211	83	215	183	Marketplace
Extra.com	25	64	143	9	15	129	152	85	50	113	Marketplace

**Table 12.** Accessibility compliance errors for non-PWA websites.

Dataset/Auditing Tools	T1 (%)	T2	T3	T4 (%)	T5	T6	T7	T8	T9	T10	Niche
Bbc.com	7	3	4	2	59	47	127	10	17	82	News Blog
Nike.com	19	14	139	31	70	224	109	31	62	116	Clothing Brand
Doordash.com	22	3	4	33	0	6	14	1	10	19	Food Ordering
Booking.com	12	2	41	14	59	15	729	5	198	99	Ticketing
Dw.com	9	4	15	19	16	25	192	36	288	308	News Blog
Usatoday.com	18	9	33	22	12	99	483	126	175	44	News Blog
Harrods.com	11	1	0	33	10	739	61	16	28	46	Luxury cosmetics
Hopper.com	89	4	8	8	0	32	131	48	57	131	Travel Website

Nbcnews.com	43	30	64	22	454	33	472	29	72	106	News Blog
Americacarrental.com	28	27	19	21	22	79	41	15	25	19	Car Rental
Gumtree.com	5	6	64	17	14	11	237	47	159	178	Free Classifieds
Facebook.com	11	3	1	0	4	22	79	27	70	92	Social Network
Expedia.com	9	3	0	22	43	108	30	7	62	57	Travel Booking
Zillow.com	8	2	2	24	11	82	20	1	7	14	Property Website
Snapdeal.com	35	227	87	7	27	309	566	431	683	276	Marketplace
Weather-forecast.com	13	4	5	5	11	21	62	7	149	125	Weather Platform
Royalcaribbean.com	21	13	3	12	113	397	133	22	91	218	Travel Booking
Dailymail.co.uk	28	50	145	20	45	348	2134	217	1067	530	Newspaper
Ebay.com	8	26	27	12	23	319	142	43	80	115	Marketplace
Amazon.com	12	39	79	14	38	199	191	54	159	168	Marketplace

The dataset of 20 PWAs and 20 conventional non-PWA websites was manually fully tested against 10 accessibility evaluation tools—auditing tools. For each website of the dataset, a manual procedure was performed, taking the website’s URL and placing it to the auditing tool’s input area. Auditing tools, in turn, accept the URL, sending a bot to the given website to perform accessibility inspections. The results returned by the auditing tools for each website were documented in excel sheets to be easily manageable by the program that will perform later the descriptive analysis. Each tool has its own metrics for measuring and displaying the results—total errors. Two of them, Google Lighthouse and PowerMapper, show the errors as a percentage of total checks. The remaining eight tools show the results as the number of the errors found. For the sake of simplicity, a unique ID key has been assigned to each auditing tool presented in the following tables: (T1) Lighthouse by Google, (T2) Wave by WebAim, (T3) Web Accessibility by Level Access, (T4) Power Mapper, (T5) CSS Validation Service by W3, (T6) Nu Html Checker by W3, (T7) AXE DevTools, (T8) Accessibility Insights Chrome, (T9) Equal Access by IBM, and (T10) Site Improve. The evaluation results for PWA websites are presented in **Table 11** and for non-PWA in **Table 12**, respectively.

To analyze our dataset to achieve valuable results, Jupyter software has been used. Jupyter Notebook is the most widely used system for interactive literate programming for data science purposes. It was designed to make data analysis easier to document, share, and reproduce [265]. In our analysis, for simplicity purposes, we have used the terms “older technology” for non-PWA websites and “new technology” for PWA websites, respectively. Jupyter software, although based on Python language, can support many other languages, such as Java and R. In our analysis, we have used Python 3 programming language, which is highly supported by Jupyter’s core.

**Figure 21** presents the Python libraries we have used for our data analysis and **Figure 22** presents the way we imported our dataset into Pandas Data Frame.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

plt.style.use('ggplot')
from matplotlib import rcParams
rcParams['figure.figsize'] = 15, 5

import seaborn as sns; sns.set_theme()
```

**Figure 21.** Import required in Python Libraries.

```
old_tech = pd.read_excel('OLD-TECHNOLOGY.xlsx')
new_tech = pd.read_excel('NEW-TECHNOLOGY.xlsx')
new_tech.head()
```

**Figure 22.** Read data into Pandas Data Frame.

The results after the descriptive analysis, both for new and older technology, are presented in **Table 13** and **Table 14**, respectively.

**Table 13.** Descriptive analysis for new technology (PWA)—new\_tech.describe().

	T1 (%)	T2	T3	T4 (%)	T5	T6	T7	T8	T9	T10
<b>count</b>	20	20	20	20	20	20	20	20	20	20
<b>mean</b>	20.30	31.15	35.75	16.85	13.40	155.35	161.55	53.95	104.30	128.85
<b>std</b>	15.21	50.72	52.72	10.59	19.10	273.16	126.23	60.97	99.86	91.38
<b>min</b>	3	1	1	0	0	2	21	0	7	30
<b>25%</b>	11	2	4.25	11.25	1	19.75	52.25	11.50	22	53.50
<b>50%</b>	13.50	7.50	9.50	17.00	5	58.50	140	33	56.50	114
<b>75%</b>	25.25	29.25	40	26	15.50	135.25	221.25	83	176	190
<b>max</b>	58	166	171	35	72	1000	450	233	290	371

**Table 14.** Descriptive analysis for older technology (non-PWA)—old\_tech.describe().

	T1 (%)	T2	T3	T4 (%)	T5	T6	T7	T8	T9	T10
<b>count</b>	20	20	20	20	20	20	20	20	20	20
<b>mean</b>	20.40	23.50	37	16.90	51.55	155.75	297.65	58.65	172.95	137.15
<b>std</b>	19.07	49.94	45.42	9.65	98.84	187.42	476.80	101.11	258.52	122.92
<b>min</b>	5	1	0	0	0	6	14	1	7	14
<b>25%</b>	9	3	3.75	11	11	24.25	61.75	9.25	49.75	54.25
<b>50%</b>	12.50	5	17	18	22.50	80.50	132	28	76	110.50
<b>75%</b>	23.50	26.25	64	22	48.50	245.25	295.75	47.25	163	170.50
<b>max</b>	89	227	145	33	454	739	2134	431	1067	530

The x-axis scale in both **Table 13** and **Table 14** presents the auditing tools undertaken to perform the accessibility checks on sample websites, while the y-axis scale presents the metrics used by the tool to perform the descriptive analysis. To clarify the results shown in **Table 13** and **Table 14** we provide some additional information below:

- The `count()` function used to count the sum of the data.
- The `mean()` function used to count the average value.
- The `std()` function used to compute the standard deviation along the specified axis.
- The 25%, 50%, and 75% are three quartile values (Q1, Q2, and Q3, respectively). The second quartile (Q2) is the median of the whole data, the first quartile (Q1) is the median of the upper half of the data, while the third quartile (Q3) is the median of the lower half of the data.
- The `max()` and `min()` functions used to present the max and min values of errors per auditing tool.

A way to start the descriptive analysis is by looking at what kind of ranges each tool has and how much they vary around their average values (mean). We can observe from the above plots that CSS Validation Service by W3 (T5) error value reduced to a significant level in new technology. The average error value in new technology is 13.40, while it is 51.55 in older technology. Similarly, in new technology, the maximum axe DevTool (T7) error value is 450, while it is 2134 in older technology, showing that PWAs are operating significantly better and that error rates in websites have been lowered to a larger extent. Both **Table 13** and **Table 14** demonstrate that conventional non-PWA websites retain more errors than PWA websites in checks performed by accessibility auditing tools. As a result, the PWA websites of the dataset follow the accessibility guidelines more strictly than conventional non-PWA websites. Examining the mean values in **Table 13** and **Table 14**, we notice that nine out of ten auditing tools return less errors on average for PWAs than non-PWA websites, concluding that websites that use the PWA technology are more compatible with the existing accessibility guidelines.

To further support our findings, we graphically present the results for each tool into plots in **Figure 23** and **Figure 24** and **Figure 25**.

```

fig,axs=plt.subplots(nrows=2)
fig.set_size_inches(15,10)

sns.barplot(y="Equal Access by IBM", x="Dataset/Auditing Tools", data=new_tech,ax=axs[0], dodge=False)
sns.barplot(y="Equal Access by IBM", x="Dataset/Auditing Tools", data=old_tech,ax=axs[1], dodge=False)

axs[0].set_title('Figure 1: Access by IBM in New-Technology')
axs[1].set_title('Figure 2: Access by IBM in Older-Technology')
axs[0].tick_params(axis="x", labels=12, rotation = 45)
axs[1].tick_params(axis="x", labels=12, rotation = 45)
plt.tight_layout()
plt.show()

```

Figure 23. Presenting outcomes into plots—Equal Access by IBM (T9).

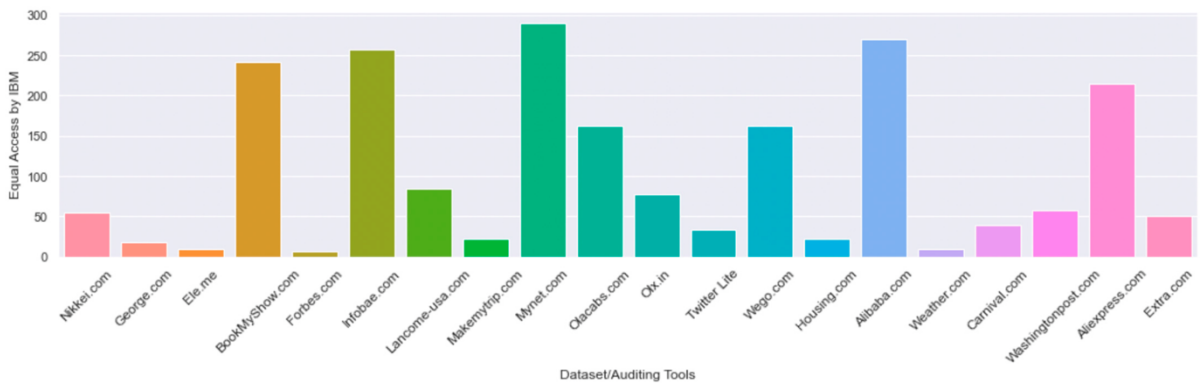


Figure 24. Equal Access by IBM (T9) errors for new technology (PWA) websites.

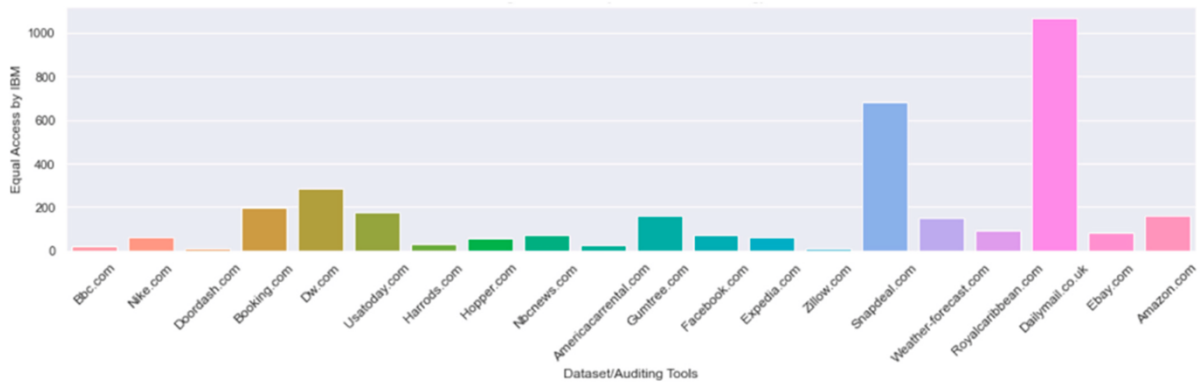


Figure 25. Equal Access by IBM (T9) errors for Older technology (non-PWA) websites.

Figure 24 and Figure 25 compare errors/bugs in the IBM Equal Access toolkit (T9). The y-axis scale shows that error values in modern technologies are substantially lower than in older technologies, as calculated by several auditing tools. The last column in both Table 11 and Table 12 presents the niche—industry for each website checked, grouping for extension of our datasets. The calculation of errors per industry is depicted in Figure 26, while the corresponding plot is presented in Figure 27, based on Axe DevTools (T7).

```

# Using plotly.express
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

fig = make_subplots(rows=1, cols=2, subplot_titles=("New technology", "Older technology"))

fig.append_trace(go.Bar(
    x=new_tech['Niche'],
    y=new_tech["axe DevTools"],
), row=1, col=1)

fig.append_trace(go.Bar(
    x=old_tech['Niche'],
    y=old_tech["axe DevTools"],
), row=1, col=2)

fig.update_layout(height=600, width=1000, title_text="Axe DevTools")
fig.show()

```

Figure 26. PWA vs. non-PWA industry plots—Axe DevTools (T7).

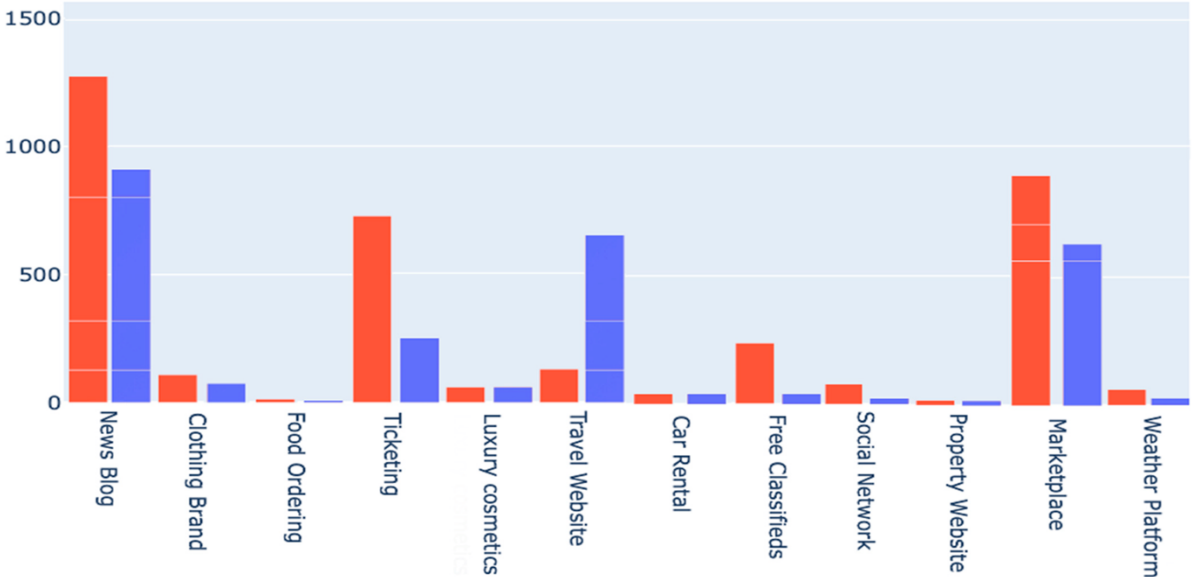


Figure 27. Presenting outcomes per industry into plots (red: older technology, blue: new technology)—Axe DevTools (T7).

The plot in **Figure 27** illustrates that the Axes Devtool (T7) error rate is reduced in PWA when compared to the corresponding non-PWA performance. As a result, we can conclude that the general performance of websites in new technology is far better than the older ones with regards to accessibility conformance to the respective guidelines. However, many of the websites still need to be improved. It is crucial to mention that the websites that theoretically have the most traffic—visitors, such as news blogs and marketplaces, are the ones that mainly ignore the accessibility guidelines, either on PWA or non-PWA websites.

To conclude with, the results confirm that each auditing tool performs its own measurements and displays its own errors based on its assessment. Measurements in Google Lighthouse (T1) indicate that PWAs emphasize performance and SEO. Although, the majority of PWAs present less accessibility errors than the non-PWAs, and only few of the former ignore in a great scale the accessibility guidelines. This observation confirms that PWAs are not fully accessible by default and that web developers must make an effort to conduct both machine and manual audits to achieve actual accessibility.

To expand our research, five ready-made, premium PWA templates from ThemeForest were selected and analyzed using the same accessibility tools. ThemeForest is a ready-made template repository where web designers and developers sell their templates. These templates state on their

purchase page that they fully adopt PWA techniques, including accessibility. However, the results from the measurements revealed that none of them follow the accessibility guidelines. The sales of these five templates have reached 2,000, meaning that 2,000 new websites can be added to the list of non-accessible websites on the World Wide Web. Template repositories are the only place where most website owners buy their templates. If these sources ignore the need for accessibility, then the majority of websites will ignore this need as well.

### ***3.6. Section Conclusion***

Web accessibility aims to provide usable web information and services to as many people as possible [266]. A growing number of technologies are characterized as accessible without actually following the accessibility guidelines. An essential need exists to adapt accessibility across to the World Wide Web, covering each platform and each technology. In this study, we conducted an in-depth review of both PWAs and web accessibility as technologies, and we combined them to evaluate PWAs' effectiveness as regards the accessibility they offer. Following specific methodology, a representative sample of 20 PWAs and 20 conventional non-PWA websites and e-commerces in corresponding niches were collected and analyzed against 10 accessibility evaluation tools. The results have shown a great lead on accessibility guidelines' conformance for the websites that have adopted the PWA architecture. We conclude that PWAs constitute a new technology with many limitations, which it exceeds daily. Despite these limitations, PWAs offer early adopters a lead in performance, SEO, and accessibility. Web developers' awareness is key to achieving the upgrade of the World Wide Web to a place where every user, regardless of his/her ability, can have equal and trouble-free access. PWAs as a technology, accessibility guidelines, and accessibility auditing tools are living organisms that evolve according to the needs of people with disabilities. Human-centered future research could create tools that will not only check websites based on a list of accessibility guidelines, but incorporate suggestions from the individuals who face a problem highlighting the website's area that is difficult for them to read, understand, or access.

## 4. Search Engine Optimization (SEO) in E-commerce

*Adapted from:*

Roumeliotis, K.I.; Tselikas, N.D. *A Machine Learning Python-Based Search Engine Optimization Audit Software*. *Informatics* 2023, 10, 68. <https://doi.org/10.3390/informatics10030068>

Roumeliotis, K.I.; Tselikas, N.D. *An Effective SEO Techniques and Technologies Guide-map*. *Journal of Web Engineering* 2022, 21, 5, 1603–1650. <https://doi.org/10.13052/jwe1540-9589.21510>

Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. *Airlines' Sustainability Study Based on Search Engine Optimization Techniques and Technologies*. *Sustainability* 2022, 14, 11225. <https://doi.org/10.3390/su141811225>

Roumeliotis, K.I.; Tselikas, N.D.; Tryfonopoulos, C. *Greek Hotels' Web Traffic: A Comparative Study Based on Search Engine Optimization Techniques and Technologies*. *Digital* 2022, 2, 379-400. <https://doi.org/10.3390/digital2030021>

Roumeliotis, K.I.; Tselikas, N.D. *Search Engine Optimization Techniques: The Story of an Old-Fashioned Website*. *International Conference on Business Intelligence and Modelling*. *Springer Proceedings in Business and Economics*. Springer, Cham 2021, 47–55. [https://doi.org/10.1007/978-3-030-57065-1\\_4](https://doi.org/10.1007/978-3-030-57065-1_4)

### 4.1. Introduction

From the very first “document-based” Web 1.0 founded in 1990, proceeded to the “social and mobile” Web 2.0 founded in 1999 and shifted to the “semantic” Web 3.0, the WEB has rapidly grown in only three decades [267]. The same rapid growth occurred in web technologies [268].

In the past, the number of websites was restricted, so the web users knew in advanced the URL of the website they wanted to visit. In current Web, with almost up to 2 billion of live websites [269], search engines used as a way for users to find the information they are looking for. Search engines act as an intermediate between websites and web users, and could be considered as a bookmark that points to specific pages depending on the query.

Web users search words or phrases, called keywords, in search engines. After each search query, users expect search engines to answer their query. The answer comes in the form of a list of websites related to the keyword and can provide the user with the answer he/she is waiting for. The results that appear after each search are called Search Engine Result Pages (SERPs). Some of the results are related to paid advertisements and some others appear strictly algorithmically without advertising payments, called organic search results. Paid results appear higher in searches and they are marked with an “Ad” badge. On the contrary, organic results are displayed below the ads. As a consequence, some website owners pay per click (PPC) to appear higher in the search results—paid results—and some others are ranking organically below the ads without any payment—organic results.

Under these circumstances, the need to optimize websites according to search engine guidelines was produced with the ultimate goal of a better position in the search results.

Search engines are constantly trying to improve their algorithms in order to provide the best possible results to user searches [270]. According to Google’s search engine, there are over 200 factors that AI algorithms take into consideration to rank websites in each search query in the organic results area [271,272]. Search engines do not publicly reveal their algorithmic methods for ranking websites [273]; as a consequence, SEO experts have to follow every available rule, making changes to the website source code and uncover its effectiveness. Search engine optimization (SEO) is the process used by SEO experts to improve both the quality and quantity of traffic to a website from search engines [274]. SEO experts have introduced SEO techniques intended to improve the ranking of a website in the search

results for given target keywords that assist website owners achieve the standards of each ranking factor [275].

The Internet era is characterized by the abundance of information, so search engines have become an integral tool for Internet users [276]. According to several surveys, 61% of global Internet users are looking for online products, while 44% of them are using search engines to search for any kind of product [277]. However, people tend to be interested only in the first pages' results, i.e., the results that are higher on searches [278]. While there are plenty of paid search marketing (PSM) ads, residing in the highest positions in search results, SEO techniques are trying to optimize the site's ranking position to increase (without cost) their organic traffic through search results [279]. The aim of SEO is twofold. On one hand, SEO targets to display the website correctly on search engines, while, on the other, it aims at meeting the needs of its visitors [280]. More specifically, the term SEO refers to all these techniques used to optimize the website appearance, the webpage code and the webpage content as well [276]. By applying SEO techniques to a website, all usability, user experience (UX) and content quality are increased resulting to higher scores in the search engine rankings, respectively [281].

SEO techniques and technologies are grouped into two main categories, on-page SEO and off-page SEO, respectively. On-page SEO techniques concern the changes and additions that the website must apply to its source code in order to achieve each ranking factor. On the contrary, off-page SEO techniques concern tasks that must be carried out on third web sources that will give credibility to the website [273].

When implementing a SEO technique, individual changes and/or modifications are made at least in some parts of the website. These kind of modifications do not often lead to the expected results, but, if combined with more SEO techniques, they could result in significant improvement as regards the organic traffic of the website [282]. The SEO techniques and technologies that will be presented in this study belong to the White Hat SEO techniques category, i.e., techniques that promote quality content and do not try to trap web crawlers, as Black Hat SEO techniques do [277,283]. Black Hat SEO techniques usually attempt to increase the ranking of malicious pages in search engine results for popular search terms [284]. Just for the reader who wants to find extra information about Black Hat SEO techniques, Content Automation [285], Cloaking [286] and Guest Posting Networks [287], are listed among the most well-known of them, but they are out of the scope of this study, since this kind of techniques are trying to violate the terms of use of search engines.

Past related research articles manage to identify what SEO techniques exist and how important they are to achieving better positions in search results. However, none of these publications suggest exactly which SEO techniques a website owner should use and in what order to optimize SEO results.

Our research came to fill this gap by examining which SEO techniques will produce better results.

The presentation starts with a detailed review of all SEO techniques from the existing literature, extending the study to SEO technologies that can help a website from a technical point of view to achieve a better presentation to both search engines and end-users who browse it.

**Table 15** contains all the SEO techniques and technologies were reviewed in this chapter.

**Table 15.** SEO techniques and technologies summary table.

SEO Techniques	Technologies Promoting SEO
4.2.1 Copywriting and Keyword Optimization	4.4.1. Responsive Design - Mobile Friendliness
4.2.2 Crawlable URLs – SEO Friendly URLs	4.4.2 Website Speed
4.2.3 Descriptive Meta Tags	4.4.3 Serving Data over HTTPS
4.2.4 Heading Tags	4.4.4 Schema.org and Structured Data
4.2.5 Optimizing Images for Search Engines	4.4.5 Accelerated Mobile Pages Project (AMP)
4.2.6 Tags on Images and URLs	4.4.6 Apache Deflate and Gzip
4.2.7 Rating and Review	4.4.7 Page Caching
4.2.8 Sitemaps and RSS Feed	4.4.8 Minify Resources (HTML, CSS, and JavaScript)
4.2.9 Robots.txt	4.4.9 Opengraph Protocol

The remainder of the **Chapter 4** is organized as follows: **Section 4.2** analyzes as a review on-page SEO techniques that optimize, in terms of presentation, a website both to its visitors as well as to search engines. These techniques and technologies give particular emphasis on the structure and the layout of the website, as well as to its content. Off-page SEO techniques are presented in **Section 4.3**. **Section 4.4** introduces the new technologies that promote SEO, focusing on the special speed and security, as well as on the implementation of structured data in the website. **Section 4.5** briefly presents the Black Hat SEO techniques. Paid Traffic is analyzed in **Section 4.6**.

During the course of the research, three web-crawler-like software programs were implemented in PHP and Python. These programs, through scraping and ML algorithms, aim to identify the SEO techniques and technologies employed by a website, as well as the strategies adopted by its competitors for specific keywords. Through thorough competitor analysis, they aim to provide valuable insights into the techniques that a webmaster should follow to secure top positions in search results, surpassing the competition, and consequently increasing organic traffic and, by extension, sales.

To confirm the effectiveness of these specific SEO software programs, individual case studies were conducted, revealing significant insights into SEO techniques followed by specific business domains under the umbrella of e-commerce, such as hotel booking platforms and airline booking platforms. As a result, in **Section 4.7**, the methodology for developing one of the three software programs in Python with the assistance of ML models is presented.

In **Section 4.8**, the first case study is introduced, involving a sample of 1,689 e-commerce websites and 311 other websites, for which an evaluation was conducted regarding the SEO techniques they employ. **Section 4.9** presents a corresponding case study for airline companies, and **Section 4.10** presents a case study for Greek hotels. As a result of these studies, the most effective SEO techniques and technologies for achieving better search rankings were identified, with a focus on competition for specific keywords.

In the final stage of the research (**Section 4.11**), the most dominant SEO techniques were utilized to upgrade an older e-commerce platform, yielding long-term results with a remarkable 782% increase in organic traffic.

## 4.2. On-Page SEO Techniques

On-page SEO techniques are related to the modifications that must be implemented in the website's source code to meet each ranking factor's requirements [281]. These techniques are designed to enhance the visitor's browsing experience (UX) by providing useful and high-quality content. In contrast, off-page SEO is a set of techniques applied outside the context of a website [273]. Off-page techniques aim to influence rankings through SERPs, primarily by increasing authority on the domain through acquiring links from other websites [273]. This section primarily focuses on the analysis of on-page SEO techniques and technologies. In **Section 4.3**, off-page SEO techniques will also be presented.

### 4.2.1. Copywriting and Keyword Optimization

Content is the most important ranking factor on a website [282]. The sustainability, the traffic and the rise of a website, depends mainly on its content [288]. The primary reason a searcher visits a website is, in most cases, to find answers to one of their problems, needs, or questions. For example, they may be looking for a book, wanting to buy a pair of shoes, searching for the address of a local business, or seeking to read the news, etc. All potential solutions to their problems, needs, or questions comprise the content of websites. Without content, websites have no reason to exist, as visitors wouldn't find answers to any of their problems, needs, or questions.

## Keyword Analysis and Research

Keywords are the words entered by a user into a search engine to find specific content on the internet (search queries). Users input keywords based on their requirements and perceptions [289]. Through searches, keywords become the crucial factor bringing visitors to a website. SEO experts are aware of the keywords that should be employed to rank their websites high. However, a keyword analysis needs to be conducted to identify a set of keywords for optimizing the website [276]. This analysis involves determining which keywords are relevant to the website and, simultaneously, are frequently searched in search engines. Google Adwords is an efficient tool for discovering the most used keywords [290]. By entering a keyword, Google Adwords suggests similar keywords and classifies them according to their corresponding frequency appearances in searches.

Keywords can be organized into levels, starting from Level 1 for more general keywords up to Level 4 for very specific ones. As the levels increase from Level 1 to Level 4, the keyword-specificity also increases, resulting in fewer results and traffic in a search, as depicted in **Table 16**. On the contrary, as the level decreases from Level 4 to Level 1, the keyword becomes more general and yields more results in a search [280]. The above analysis concerns the query of a user in search engines, while, from the website's perspective, the more specific a keyword is, the easier to rank the website higher in searches is, as well. Typically, the so-called long-tail or medium-tail keyword increases the likelihood of finding the site higher in search results [282].

**Table 16.** Keyword Level Types

Level	Keyword	Avg. monthly searches
Level 1	Apple	3.350.000
Level 2	Apple macbook (Medium-Tail Keyword)	49.500
Level 3	Apple macbook pro (Long-Tail Keyword)	40.000
Level 4	Apple macbook pro m3 (Long-Tail Keyword)	10.000

## Descriptive Title Elements

According to the W3C, the title tag is a fundamental part of a website. The title consists of a combination of words and phrases that represent the content of the web page [291]. As long as the title is descriptive and concise, it is used by search engines to add the website to the search results [271]. Its relevance to the search query gives the searchers a quick insight into the content as far as they decide which of the results to click.

The title tag should have a specific length size to positively affect both web users' experience and search engines. The recommended title length should be neither less than 6 nor more than 78 characters [292]. There are many cases where the title tag is misleading for ranking purposes, including keyword stuffing techniques. In the corresponding cases, search engines replace the search result title with a tag considered more relevant to the content of the web page [271]. As reported by Moz's testing and experience, keywords closer to the beginning of the title tag may have more impact on search rankings [293].

## Articles and Content

The websites consist of text, images and multimedia (audio/video). Text can be articles, business information, terms of use, etc. Every text within a website should be optimally structured, but not overoptimized, to attract not only visitors, but also search engines [294]. The Google search engine in Search Console Help has announced that the site should be primarily human first [295]. For this reason, the text of a website should have as its primary purpose to be useful to its visitors [296]. Once a text that meets the needs of the visitor is created, it can then be optimized to make it look better for search engines, too. When search engines parse a text or article they are not really aware of the actual content and what exactly it is dealing with. To fill this gap, a search engine reads the title of the article, meta tags, and content using advanced AI models to understand the subject matter of the article. A

digital marketing technique that deals with the process of creating a friendly text to both visitors and search engines is SEO copywriting, based on a set of rules that content's structure should follow, as shown in **Table 17** [297].

**Table 17** presents the essential guidelines for SEO Copywriting and explains how to effectively display keywords on a single page of the website. As far as the content is concerned, another concept that must be taken into account is keyword density, i.e., the frequency that a keyword displays between the text of a page. According to [292] and [276], the percentage of occurrence of the keyword within a page should be between 3% and 8% of the total number of words. Also, keyword distribution is the distribution of the keyword in the structure of an HTML document to deliver the best possible search results. A keyword can be placed in the page title, the meta keywords section, the meta description section, the page header tags (H1,..., H6), the internal links, the external links, the alt attribute of the images, and the URL [276]. It should be noted that the over-optimization of the page's content should be avoided, because it may produce negative results, since the most important aspect is to create unique content or articles that fulfill the searcher's goal and address the problem that the searcher had when accessing the website [276].

**Table 17.** Copywriting Structure

Rule Title	Copywriting rules
<b>Page Title</b>	<ul style="list-style-type: none"> <li>▪ The recommended title length is between 6 and 78 characters [292].</li> <li>▪ The target keyword should be included in Page Title and would be beneficial if used at the beginning of this [276,298].</li> </ul>
<b>Page Description and Meta Description</b>	<ul style="list-style-type: none"> <li>▪ Meta Description must be at least 150 characters long and should not be longer than 312 characters [299].</li> <li>▪ The target keyword should be included in Page and Meta Description [276,298].</li> </ul>
<b>Main Content</b>	<ul style="list-style-type: none"> <li>▪ Main Content should consist of at least 500 words [300].</li> <li>▪ The target keyword should be within the first 50 words of Content [300].</li> <li>▪ The recommended Keyword Density is between 2% and 8% [292,298].</li> <li>▪ Content should be hierarchically structured correctly using h1, h2, and h3 headings [298].</li> <li>▪ In the headings, it would be useful to add the target keyword - Keyword Distribution [276,298].</li> <li>▪ It would be useful for the target keyword and its synonyms to appear in bold and italic format where they are used in the text [298,300].</li> <li>▪ Internal links should be added to refer to the website itself [298], parsing authority from one page to another.</li> <li>▪ External links should be added which might explain the meaning of a word [298].</li> <li>▪ Content must contain at least one image with SEO friendly URL and SEO friendly alt tags [300].</li> </ul>

#### 4.2.2. Crawlable URLs – SEO Friendly URLs

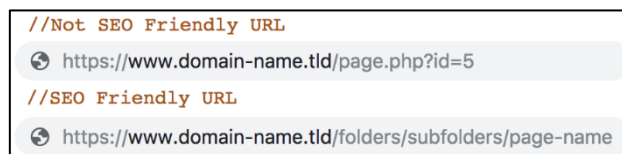
Web crawlers work persistently in the background by following links to collect updated data from the World Wide Web and store them in large databases, which are used by search engines, like Google, to fulfil the search queries of searchers [298]. Indexing by a search engine is the process of creating an indexed list, consisting of websites and their associated content [298]. Once a search request occurs, the search engine processes it and compares it to the data stored in its database. Finally, search engine algorithms undertake to display the most relevant results and present them to the user [283].

URL (Representational state transfer (RESTful) URLs or search-friendly URLs or user-friendly URLs) is a human-readable text that defines the structure of files within a web server. Each URL has three distinct parts and consists of the access protocol, the domain name, and a path [301]. The URLs of a website are crucial ranking factors and among the fundamental elements that search engines consider to comprehend a website's content and, consequently, connect it to a search query or not. A well-structured URL provides both visitors and search engines an easy way to understand the page's content even before visiting it. Search engines also expect the target keyword to be present in the URL keywords [302].

Additionally, several times, users copy a URL to share it via email or through social media. If the URL is well written, legible and understandable, it is more likely that users will click on it [302].

#### Differences between SEO URL & non SEO URL

Years earlier, most URLs were illegible, making it difficult for search engines to understand the content of the page. On the other side, SEO Friendly URLs contain the page path using words that describe the content, separated by hyphens (-), as depicted in **Figure 28** [303]. A common practice is to use the title tag of a page as a URL, since it usually contains keywords related to the search query [282].



**Figure 28.** Differences between SEO friendly and Not SEO friendly URL

#### How SEO Friendly URLs are Created

In the not SEO friendly URL case, the corresponding URL syntax depends on the server-side scripting language. For example, by typing the above URL (**Figure 28**), the `id=5` clause implies an article having the value five as the unique primary key in a database's table, but there is nothing clear about its content. On the other hand, to create SEO-friendly URLs the developer has to put an effort using a scripting language in a combination with a web server software. There are multiple ways to achieve SEO-friendly URLs. Two of them are:

- i. Many opensource CMS platforms, like Wordpress, use the front controller design pattern. The Front Controller pattern consists of a central controller that serves as a single point of entry into the application. Either using Apache, NGNIX or ISS web server software, the developer has to redirect all the traffic to a single point of entry using the appropriate URL Rewrite Module [304]. In this single point of entry, regardless of the programming language used, the front controller undertakes to break the given URL into sections that can help locating the content in the database.
- ii. An alternative way to achieve SEO-friendly URLs is to use the prior URL Rewrite Module to rewrite every single type of URL pointing to the appropriate scripting file. A rewrite rule using the `.htaccess` on Apache and PHP programming language is presented in **Figure 29**. In this example the rewrite rule matches the URL provided by the browser, identifies the appropriate rule, and redirects the user to the `blog.php` file with a `HTTP GET` request.

```
Options +FollowSymLinks -MultiViews
RewriteEngine on
RewriteBase /
RewriteCond %{REQUEST_URI} !\.(?:css|js|jpeg|gif|png|ico|eot|ttf|otf|woff|html|php)$ [NC]
RewriteRule ^blog/([A-Za-z0-9-]+)/?$ blog.php?article=$1 [NC,L]
```

Figure 29. Source code from an .htaccess file

At this point, it's important to note that URLs should not exceed 2083 characters, in order to ensure visibility on all browsers, including mobile searches [282,305,306].

#### 4.2.3. Descriptive Meta Tags

The meta description tag follows the title tag in the head container [291]. When searching and reviewing search results, users typically see the page's title first, followed by the URL, and finally the meta description. This HTML tag serves as a concise abstract for the page, presenting real sentences related to its content [275]. The meta description provides users with information about what they are about to find on the page [273]. SEO experts optimize meta descriptions to ensure that listings are compelling and consistent with the website's content [305]. It should be at least 51 characters long but not exceed 350 characters [20]. Including the target keyword in the meta description can also enhance SEO results [276].

Results indicate that pages containing metadata tend to rank higher than those without [274]. However, it's noteworthy that, despite a webpage having a meta description tag, the search result description may differ. Search engines, mostly Google, often consider meta descriptions as non-representative, because they can be misleading. This is due to instances where the <meta> description is not correctly crafted to describe the exact content of the page. Advanced search engine algorithms determine whether <meta> description or <meta> keywords are relevant and may replace them with text from the page deemed more pertinent.

It is advisable to create a <meta> description for each page that accurately describes the page content accurately in 50 to 160 characters [307]. For optimal visibility in search queries, the <meta> description could contain the target keywords, as depicted in **Figure 30**.

```
<head>
<meta charset="UTF-8">
<title>Macbook Pro MJLQ2</title>
<meta name="description" content="Macbook Pro, 2.2GHz Intel Core i7 processor with
Turbo Boost up to 3.4GHz, the MacBook Pro delivers speedy performance.">
<meta name="keywords" content="macbook,apple macbook">
</head>
```

Figure 30. Appropriate <meta> description and <meta> keywords

#### 4.2.4. Heading Tags

According to w3c web standards, HTML defines six levels of heading tags. The heading elements are H1, H2, H3, H4, H5, and H6, with H1 being the highest (or most important) level and H6 the least [308]. The Google Central Blog urges webmasters to use heading tags to emphasize the important parts of a text [271]. Heading labels are practically the highlights of a site. Heading names are given more worth by means of web indexes than ordinary body copywriting [291]. It can also be a great pathway used by screen readers for people with visual disabilities. In addition, H1 and H2 tags are important when a web crawler examines a website. For SEO purposes, the main keyword should appear in both H1 headings and H2 subheadings and the length should be between 10 and 13 words [309]. In SERPs, the title tag could be ignored by the search engines and, in these cases, it is replaced by the H1 one.

#### 4.2.5. Optimizing Images for Search Engines

Each page, whether containing an article or a text describing the activity of a business, includes many times images as well. Images contribute to better ranking in search engines, because they can enhance the user experience, provided that they are in the proper form, i.e., with an appropriate file size first [310]. Independent studies indicate that images exceeding 100 kb in file size are often

challenging for users to view, despite recent improvements in internet speeds [310,311]. Even today, many pages are slow to load, which prevents users from visiting them. Thus, an essential factor influencing a website's load time is undoubtedly the file size of the images displayed on it [277]. Utilizing Next-Gen formats such as JPEG 2000, JPEG XR, and WebP, along with an Image CDN can significantly increase a website's speed [310,312].

#### 4.2.6. Tags on Images and URLs

In addition to the file size of an image, the file name and alt tags play a crucial role in SEO. Although algorithms have evolved by applying ML techniques to identify the content of images, alt tags and filenames are considered more representative and indicative of the description of images on a website [313]. Apart from searching for websites, many search engine users choose to search for images in order to decide which website fits their needs. For both reasons, the file name should be relative to what the image is displaying, and the alt tag should describe what the image represents, too. Because the images are part of a page and are related to the content of the page, it would also be beneficial to contain the page title (**Figure 31**). Alternative tags also contribute to better accessibility. Starting from the WCAG 2.0 the W3C has included alternative tags in the accessibility guidelines. According to the success criterion 1.1.1 (Non-text Content), when using the `img` element, it is mandatory to include a short text alternative with the `alt` attribute [314]. Screen readers report alternative text in place of images, helping users with visual or certain cognitive disabilities perceive the content and function of the images [315].

```
<!-- Not SEO Friendly Image Tag -->  
  
  
<!-- SEO Friendly Image Tag -->  

```

Figure 31. Image SEO Filename & Alt Tag

Each page contains internal and external links. A link could refer, for example, to an article or a product. When a user sees a link on a page, they usually hover the mouse over it clicking. Before clicking this link, the user would probably want to know where that link goes. Similarly, search engines also want to know about the content of the landing page [305]. For all the above reasons mentioned above, the title tag is necessary for both describing and the understanding of an internal or external link (**Figure 32**).

```
<!-- Not SEO Friendly URL Tag -->  
<a href="/product.php?id=2">  
  
<!-- SEO Friendly URL Tag -->  
<a href="/product/macbook-pro-mj1q2/" title="Macbook Pro MJLQ2">
```

Figure 32. Link SEO & Title Tag

#### 4.2.7. Rating and Review

Many users search for products and services on the Internet, and a significant number of them ultimately make a purchase on the corresponding ratings and/or reviews. User-generated content (UGC) is an effective tool for selling online products and services, as consumers often decide to purchase an item after reading personal information generated by other users on the platform [316]. The psychological reaction that leads users to believe the reviews of other users of a website are accurate is called social proof [317]. From the side of the website or e-commerce, social proof can significantly increase conversion rates by building a feeling of customer trust [318].

As a result of the above, by implementing a rating and review system, each website can benefit from user ratings and reviews. Through this system, users create valuable content for the website while encouraging other users to trust their choices. Search engines, in turn, reward websites that promote

user ratings and reviews., e.g., the Google search engine often prominently displays user site reviews in search results (Rich Results) (Figure 34).

In order for search engines to be in position to locate the rating system, the website should contain structured data describing the reviews (Figure 33).

Structured data can be written in three ways, i.e., with Microdata that are the widely used structured data, with Resource Description Framework in attributes (RDFa) or with JavaScript Object Notation for Linked Data (JSON-LD) [319]. Microdata is a set of Web Hypertext Application Technology Working Group (WHATWG) HyperText Markup Language (HTML) specification properties that are used to display metadata on existing page content. Search engines, web crawlers and browsers extract and process microdata from a website and use them to provide a richer browsing experience for users [320]. RDFa is a W3C recommendation that adds a set of attribute-level extensions to existing HTML and XHTML to import metadata into web documents [321]. JSON-LD is a metadata encoding method using JSON [322]. Microdata and RDFa are applied directly to HTML by adding HTML properties. According to W3C use cases, JSON-LD has many advantages over RDFa since JSON-LD templates can be easily inserted as a single block of structured data within the <head> element of a web page [323].

```

A
<!-- MicroData Rating and Review -->
<div itemscope itemtype="http://schema.org/Review">
  <div itemprop="itemReviewed" itemscope itemtype="http://schema.org/Product">
    
    <span itemprop="name">Macbook Pro MJLQ2</span>
  </div>
  <span itemprop="reviewRating" itemscope itemtype="http://schema.org/Rating">
    <span itemprop="ratingValue">4</span>
  </span> stars -
  <b><span itemprop="name">A good product.</span></b>
  <span itemprop="author" itemscope itemtype="http://schema.org/Person">
    <span itemprop="name">Bob Smith</span>
  </span>
  <span itemprop="reviewBody">The product is great.</span>
  <div itemprop="publisher" itemscope itemtype="http://schema.org/Organization">
    <meta itemprop="name" content="Washington Times">
  </div>
</div>
B
<!-- JSON-LD Rating and Review -->
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "Review",
  "itemReviewed": {
    "@type": "Product",
    "image": "https://www.website.com/apple-macbook-mjlq2.jpg",
    "name": "Macbook Pro MJLQ2"
  },
  "reviewRating": {
    "@type": "Rating",
    "ratingValue": "4"
  },
  "name": "A good product.",
  "author": {
    "@type": "Person",
    "name": "Bob Smith"
  },
  "reviewBody": "The product is great.",
  "publisher": {
    "@type": "Organization",
    "name": "Washington Times"
  }
}
</script>

```

Figure 33. Structured Data Ratings and Reviews – A) Microdata and B) JSON-LD

Amazon.com: Apple 15 Inch MacBook Pro Laptop (Retina Display, 2.2 ...  
<https://www.amazon.com/Apple-MacBook-Display...MJLQ2LL/dp/B00XZGMBVC> ▼  
 ★★★★★ Rating: 4.3 - 296 reviews  
 PERFORMANCE: 2.2GHz Intel Core i7 processor with Turbo Boost up to 3.4GHz, the MacBook Pro delivers speedy performance. Includes 16 GB of memory so ...  
 Apple MacBook Pro MJLT2LL · Apple MacBook Pro MJLQ2LL ...

Figure 34. Ratings, Reviews and Feedback on Search Results

#### 4.2.8. Sitemaps and RSS Feed

The sitemap is an XML file that provides information about the pages of a website [324]. By reading this file, search engines can better map the layout of a website. Each line in the sitemap describes to search engines, not only the structure of the website and which pages are more important than others, but also how often the information is updated (Figure 35). Without the sitemap, search engines would need to check the internal linking of the website to determine the layout, risking a misunderstanding and incorrect indexing.

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.website.com/article-url</loc>

    <lastmod>2018-12-01</lastmod>

    <changefreq>monthly</changefreq>

    <priority>0.8</priority>
  </url>
</urlset>
```

Figure 35. Sitemap Example

The RSS feed is also a decisive factor in improving the visibility of a website by search engines. The RSS format was originally developed by Netscape in the late 1990s for use on My Netscape Network, a customizable start page for Netcenter [325]. My Netscape Network is provided a simple RSS framework for websites to create channels that can then be added to the customizable start page [326]. Currently, the primary use of the RSS feed for websites is to inform visitors about the new data added to a website [327]. Typically, blog sites use RSS to offer their users a more immediate update. To enhance users' experience, RSS feed readers plugins have been created to help users stay informed about developments in topics that interest them. From the search engines' perspective, they are more likely to locate the most recent content on a website by following its RSS feed, since it contains the information properly structured to be read programmatically by a search engine [277], (Figure 36).

```
<rss version="2.0">
<channel>
  <title>Website RSS feed</title>
  <link>https://www.websites.com</link>
  <description>The latest content of Website</description>
  <language>en-us</language>
  <copyright>Copyright (C) 2018 website.gr</copyright>
  <item>
    <title>Article Name</title>
    <description>Article Description</description>
    <link>
      https://www.website.com/article-url/
    </link>
    <pubDate>Sat, 01 Dec 2018 19:23:31 +0000</pubDate>
  </item>
</channel>
</rss>
```

Figure 36. RSS Article Example

#### 4.2.9. Robots.txt

The Robots.txt file is a text file created by webmasters to guide web robots on how to crawl the pages of a website [277]. Web robots, follow links, and visit millions of websites daily, engaging in spidering. When a robot arrives at a website, its first action is to look for the robots.txt file. In case such a file is not available, the robot investigates the content of the site by crawling into it [328]. However, if the robots.txt file exists, it first read it and then enters as many pages it has access to. In Joomla CMS, web administrators can use the robots.txt file to allow access to web content while simultaneously

prohibiting access to pages that they do not want to be indexed, such as the administrator panel [329] (Figure 37).

```
#Robots.txt on public_html                                #Robots Meta
User-agent: *                                             <html>
Allow: /                                                  <head>
Disallow: /administrator/                               <title>Administrator</title>
Sitemap: https://www.websites.com/sitemap.xml           <meta name="robots" content="noindex, nofollow">
                                                         </head>
```

Figure 37. Robots.txt and Robot Meta

#### 4.2.10. Breadcrumbs

Breadcrumbs are text that consist of links (usually internal) and are typically located at the top of the website as shown in Figure 38 [330]. They show the visitor exactly where they are within a website. These links describe the path that a visitor should follow, for example, to reach a product. By clicking on one of the links, the visitor can return to the appropriate location.

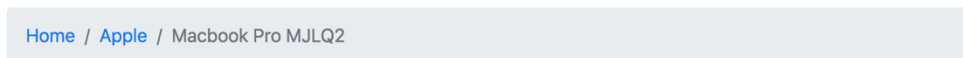


Figure 38. Breadcrumbs

In terms of visual design, Figure 39 provides an example of how breadcrumbs are created.

```
<!-- Breadcrumbs Visualization Bootstrap -->
<nav aria-label="breadcrumb">
  <ol class="breadcrumb">
    <li class="breadcrumb-item"><a href="/">Home</a></li>
    <li class="breadcrumb-item"><a href="/Apple/">Apple</a></li>
    <li class="breadcrumb-item active" aria-current="page">Macbook Pro MJLQ2</li>
  </ol>
</nav>
```

Figure 39. Using Bootstrap to Visual Design Breadcrumbs

Structured data (Microdata, RDFa or JSON-LD) can be employed to maximize the utility of Breadcrumbs, aiding users and search engines in understanding and exploring a site effectively [331], as illustrated in Figure 40.

```

A
<!-- Breadcrumbs with MicroData -->
<ol itemscope itemtype="http://schema.org/BreadcrumbList">
  <li itemprop="itemListElement" itemscope
    itemtype="http://schema.org/ListItem">
    <a itemtype="http://schema.org/Thing"
      itemprop="item" href="https://website.com/">
      <span itemprop="name">Home</span></a>
      <meta itemprop="position" content="1" />
    </li>
  <li itemprop="itemListElement" itemscope
    itemtype="http://schema.org/ListItem">
    <a itemtype="http://schema.org/Thing"
      itemprop="item" href="https://website.com/Apple/">
      <span itemprop="name">Apple</span></a>
      <meta itemprop="position" content="2" />
    </li>
  <li itemprop="itemListElement" itemscope
    itemtype="http://schema.org/ListItem">
    <a itemtype="http://schema.org/Thing"
      itemprop="item" href="https://website.com/Apple/macbook-pro-mjlq2/">
      <span itemprop="name">Macbook Pro MJLQ2</span></a>
      <meta itemprop="position" content="3" />
    </li>
</ol>

B
<!-- Breadcrumbs with RDFa -->
<ol vocab="http://schema.org/" typeof="BreadcrumbList">
  <li property="itemListElement" typeof="ListItem">
    <a property="item" typeof="WebPage" href="https://website.com/">
      <span property="name">Home</span></a>
      <meta property="position" content="1">
    </li>
  <li property="itemListElement" typeof="ListItem">
    <a property="item" typeof="WebPage" href="https://website.com/Apple/">
      <span property="name">Apple</span></a>
      <meta property="position" content="2">
    </li>
  <li property="itemListElement" typeof="ListItem">
    <a property="item" typeof="WebPage" href="https://website.com/Apple/macbook-
    pro-mjlq2/">
      <span property="name">Macbook Pro MJLQ2</span></a>
      <meta property="position" content="3">
    </li>
</ol>

C
<!-- Breadcrumbs with JSON-LD -->
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "BreadcrumbList",
  "itemListElement": [{
    "@type": "ListItem",
    "position": 1,
    "name": "Home",
    "item": "https://website.com/"
  },{
    "@type": "ListItem",
    "position": 2,
    "name": "Apple",
    "item": "https://website.com/Apple/"
  },{
    "@type": "ListItem",
    "position": 3,
    "name": "Macbook Pro MJLQ2",
    "item": "https://website.com/Apple/macbook-pro-mjlq2/"
  }]
}
</script>

```

Figure 40. Structured Data – A) Microdata, B) RDFa and C) JSON-LD to improve Breadcrumbs

Visitors often find themselves lost on a website. Breadcrumbs utilize links for the primary paths of the page, assisting visitors in finding the right path for what they are looking for and, by extension improving the overall user experience [330]. For instance, some users may land on a product page through search engine results. If the visitor wants to navigate to the category to which the product belongs, they would have to search through a vast menu of categories to find the one that the product belongs to. Breadcrumbs offer a solution to this problem by presenting the visitor the exact location of the product on a website, including the category or subcategory to which it belongs.

Furthermore, Breadcrumbs, besides aiding visitors, provide search engines with insight into how a website is structured [276]. The Google search engine rewards the use of Breadcrumbs by displaying them in rich search results, as shown in Figure 41.

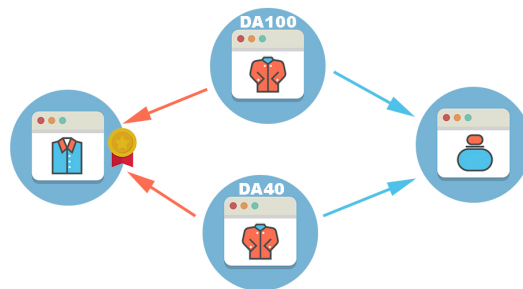


Figure 41. Breadcrumb shows website hierarchy and categorization.

### 4.3. Off-Page SEO Techniques

A backlink is created when an external link of a referring website points to a page on a receiving website. We assume that the receiving website contains a critical article while other websites encourage their visitors through a link to read the article. In this scenario, all these links are considered backlinks. For search engines, when a website has many backlinks, it is deemed reliable because numerous other websites trust it and recommend their visitors follow the links on it. Nevertheless, not all backlinks carry the same weight. A backlink from the New York Times certainly does not have the same influence as a backlink from a local business. SEO Experts rank websites according to their importance by using Domain Authority (DA) classification fluctuating between DA1 and DA100 (where DA100 characterizes highly important websites) [332].

Therefore, if a website has many backlinks with a high DA, search engines are likely to reward it by elevating its position in their search results. It is imperative to clarify that search engines don't use the DA metric to rank web pages in search results. DA is a simulated metric created by SEO experts to model the algorithms used by search engines. Another crucial ranking factor in addition to the DA of the referring website is the relevance of the content between the referring and the receiving websites [282]. If the content of these two websites is relevant, the search engines evaluate it positively, something that does not happen in the case that the referring and the receiving websites consist of irrelevant content, as shown in the example of **Figure 42**.



**Figure 42.** DA100 and DA40 Clothing Websites Point Backlink to Related Clothing Website and Not Related Perfume Website

There are various approaches to earning backlinks.

- Guest Posting or Guest Blogging Backlinks: This involves writing and submitting articles to relevant niche blog websites [273]. These blogs accept valuable articles, which in turn allow the author to place a backlink pointing to their website.
- Article Submission Backlinks: Similar to Guest Posting these backlinks are created by submitting articles, but in this case, to article directories instead of blog website [333].
- Directory Submission or Link Directory Backlinks: Created by submitting a website's name, description, and URL to popular link directory submission sites [273].
- Profile Backlinks: Implemented by creating profiles to social media platforms. The majority of the social media platforms accept a user's URL during the registration of their profile [333].
- Social Media Sharing Backlinks: Created by social media users when posting a URL of their own website to their social media profiles [273,333].
- Forum Posting Backlinks: Created while producing and commenting on discussion topics in popular forums [333]. Many forums allow users to redirect forums' traffic using URL to external resources as long as the source provides essential information solving a problem [333].
- Comment Backlinks: Created by commenting on an article or product page, placing an external link within the comment. Comments on a website are equally beneficial to search engines and visitors as mentioned in the Rating & Review (**Section 4.2.7**). However, it's important to note that earlier SEO experts attempted link building through spam comments with links on blog pages. Advanced search engine algorithms can now detect spam comments, resulting in penalties for both the spammer's website and the spammed website [273].

- Question-and-Answer (Q & A) Comment Backlinks: On many Q & A websites, like StackOverflow and Quora, there are typically three kind of visitors: (a) users seeking help on a specific subject, (b) users providing answers to questions, and (c) visitors finding solution to problems through search engines. Backlinks from such websites are beneficial. However, to create a backlink to them, the webmaster will first have to answer some of the user questions. By answering a question, the webmaster has the option to add a link pointing to their website proposing a solution to a specific problem. This added link is a backlink, which not only receives positive evaluations from search engines but also drives traffic to the website.

## 4.4. Technologies Promoting SEO

### 4.4.1. Responsive Design - Mobile Friendliness

Currently, about half of Internet users' searches originate from mobile smartphones and tablets [334]. When a user visits a website and finds it challenging to read the provided information, they are likely to abandon it quickly. Utilizing responsive design technologies enables a website to enhance access, readability, and, more broadly, improve the user experience (UX) of its visitors [277,335]. By improving these aspects, visitors spend more time on the website, leading to a decrease in the bounce rate and an increase in the average time spent on the website (Session Duration) (Figure 43).

Session duration and bounce rate are two of the most critical criteria that the Google RankBrain algorithm uses to rank websites in search engine results [288]. If a website is not mobile-friendly, it may hinder visitors from reading it, prompting them to leave quickly, thereby reducing session duration and increasing the bounce rate.

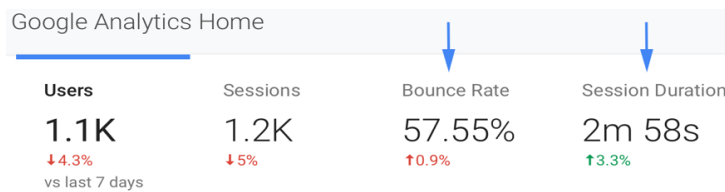


Figure 43. Google Analytics - Bounce Rate - Session Duration

In the early stages of responsive design, mobile pages were created as a separate section of the main website. The website utilized the user's HTTP headers to identify the device they used to access the website. If the user accessed the website from a mobile device, the mobile version of the website was delivered from the webserver. Conversely, when the user accessed the page from a desktop computer, the website sent the main version, using the dynamic serving technique [336]. Responsive web design [337] was developed, to address the issue of having separate websites. One of the leading companies in responsive web design, providing ready-made templates for websites viewed on various devices, is Twitter's Bootstrap [337] (Figure 44).

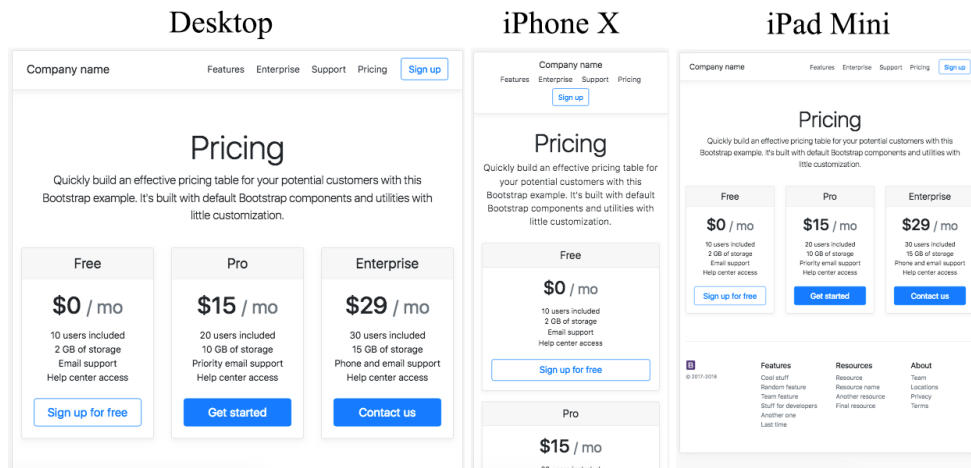


Figure 44. Responsive Design for Multiple Devices – Bootstrap

Google has created a tool called Mobile-Friendly Test. This tool checks whether the website appears appropriately on mobile devices and, if not, it informs the developer about the encountered problems [338].

#### 4.4.2. Website Speed

Webpage loading speed is a crucial factor for search engine ranking and plays a vital role in achieving better results in the SERPs [339]. Search engines consider page speed as one of the key elements a website should possess. Numerous tools are available for website owners and visitors to conduct tests on websites, including Google Lighthouse, Pingdom, GTMetrix, and SiteAnalyzer [340].

Both Lighthouse and PageSpeed Insights tools, developed by Google, assist webmasters in controlling and improving website performance. They collect data from a website and generate a performance score along with estimated opportunities for savings. The main difference between these tools is that the former only uses lab data to determine website performance on a single device and a fixed set of network conditions, while the latter employs both lab and field data to evaluate website performance on a various devices and real-world conditions. Depending on the performance score, the webmaster can proceed to recommended changes in order to achieve better performance on their website [341]. The Pingdom tool utilizes similar techniques [342].

#### 4.4.3. Serving Data over HTTPS

Every day, millions of users employ the web to carry out essential tasks, e.g., banking transactions, capital investment in the stock market, etc. Every user undoubtedly desires the interaction with a website to be secure, just as their personal data.

HTTPS encodes and encrypts the connection between the user's browser and the web server hosting the website, ensuring that data exchanged between them is not interrupted or intercepted. Any website that aims to encrypt traffic requires an SSL Certificate.

SSL Certificate can be installed on any web server to secure communication between visitors and the web server.

In 2014 Google search engine on its Search Central Blog announced that security is a top priority for Google [343]. As a search engine that redirects searchers to websites, Google wants to ensure that websites people access from Google are secure [343]. The same document provides detailed best practices to protect any website. Meanwhile, a warning from Google takes place cited that secure, encrypted connections are a signal for its search ranking algorithms, meaning that SSL certificates affect a website's ranking on SERPs.

Google Chrome version 68 was the first to highlight websites without SSL certificates as 'not secure' [344]. Nowadays, the majority of browser companies have incorporated user warnings on their user interface (UI) to protect their users from non-SSL encrypted websites. The HTTPS or HTTP/2

Internet Protocol utilizes different technologies to make a website secure and faster, thus enhancing user experience (UX). For all the above reasons the use of an SSL certificate is considered of vital importance.

#### 4.4.4. Structured Data and Rich Snippets

Schema.org is a collaborative community that creates, maintains, and promotes schemas for structured data on the Internet, on web pages, in email messages, and beyond [319,345]. Structured data are a standardized format for providing information about a page and classifying the page content [271]. Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata, and JSON-LD [18]. Schema markups can be added to the HTML source code of a website through semantic annotations that can be processed by search engines to understand the meaning of the content fragments and provide users with enriched information in the search results [320]. The use of markup formats is an emerging trend in the semantic web, in particular, because of the support of the main search engines [346]. Google rewards the use of structured data by displaying rich results—rich snippets in SERPs, such as the price and the availability of a product [20].

By utilizing schemas' the content becomes more understandable to search engines, providing them information on what the page is about. The most used item types include Organization, Person, Place, LocalBusiness, Product, and Offer. Each item type contains numerous properties. For instance, in the case of the Offer item type, these properties include the name of the offer (itemprop = name), the image of the offer (itemprop = image), the color of the offer object (itemprop = color) as well as other relevant properties.

Search engine bots browse every website and attempt to understand what each page presents. Thus, without the existence of structured data, from the phrase "Apple Macbook MJLQ2" a bot might assume that it is the title of the page, while from the "apple-macbook-mjlq2.jpg" it may infer an image illustrating the content of the page. In contrast, with structured data, search engine bots recognizes that the "Apple Macbook MJLQ2" is the name of a product, and "apple-macbook-mjlq2.jpg" is a product photo [319]. **Figure 45** demonstrates an example showcasing the use of three alternatives for structured data, i.e. Microdata (A), RDFa (B), or JSON-LD (C), respectively.

```

A
<!-- Product with MicroData -->
<div itemscope itemtype="http://schema.org/Product">
  <span itemprop="brand">Apple</span>
  <span itemprop="name">Macbook Pro MJLQ2</span>
  
  <span itemprop="description">The Apple MJLQ2 MacBook Pro with Retina
Display has a compact and suave design that accommodates immense power.
It has sleek dimensions and a narrow body that measures only 1.8cm in
height, which makes it easily portable.</span>
  <span itemprop="offers" itemscope itemtype="http://schema.org/Offer">
    <meta itemprop="priceCurrency" content="USD" />
    <span itemprop="price">2299.99</span>
    Condition: <link itemprop="itemCondition"
href="http://schema.org/NewCondition"/>New Condition
    <link itemprop="availability" href="http://schema.org/InStock"/>In
stock! Order now!
  </span>
</div>

B
<!-- Product with RDFa -->
<div vocab="http://schema.org/" typeof="Product">
  <span property="brand">Apple</span>
  <span property="name">Macbook Pro MJLQ2</span>
  
  <span property="description">The Apple MJLQ2 MacBook Pro with Retina
Display has a compact and suave design that accommodates immense power.
It has sleek dimensions and a narrow body that measures only 1.8cm in
height, which makes it easily portable.</span>
  <span property="offers" typeof="Offer">
    <meta property="priceCurrency" content="USD" />
    <span property="price">2299.99</span>
    Condition: <link property="itemCondition"
href="http://schema.org/NewCondition"/>New Condition
    <link property="availability" href="http://schema.org/InStock"/>In
stock! Order now!
  </span>
</div>

C
<!-- Product with JSON-LD -->
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "Product",
  "name": "Macbook Pro MJLQ2",
  "image": ["https://www.website.com/img/macbook-pro-mjlq2.jpg"],
  "description": "The Apple MJLQ2 MacBook Pro with Retina Display has a
compact and suave design that accommodates immense power. It has sleek
dimensions and a narrow body that measures only 1.8cm in height, which
makes it easily portable.",
  "brand": {
    "@type": "Thing",
    "name": "Apple"
  },
  "offers": {
    "@type": "Offer",
    "priceCurrency": "USD",
    "price": "2299.99",
    "itemCondition": "http://schema.org/NewCondition",
    "availability": "http://schema.org/InStock"
  }
}
</script>

```

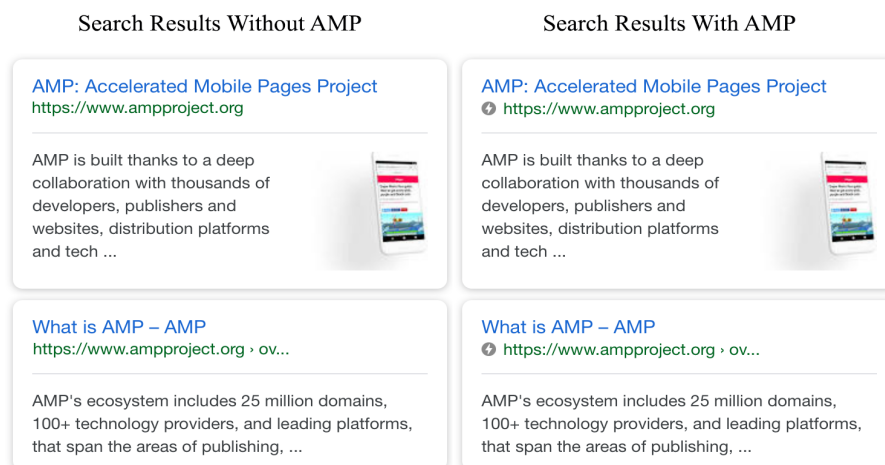
Figure 45. Structured Data on Product A) Microdata B) RDFa C) JSON-LD

In Figure 45, it is evident that applying structured data with Microdata or RDFa requires improvement to each line of the page by adding the property - "itemprop". On the contrary, with JSON-LD, all structured data is within a script element located at the top of the page, eliminating the need to correct the entire page code. To conclude, the most well-known structured data, i.e., Microdata, offers optimal understanding of the website content by search engines and is considered a valuable aid for SEO [333].

#### 4.4.5. Accelerated Mobile Pages Project (AMP)

Over the last decade, the number of mobile users has grown rapidly, surpassing 7.7 billion by late 2017 [27]. Accelerated Mobile Project (AMP) is a recent effort started by Google with a similar goal of improving the mobile browsing experience [22]. Google identifies which pages follow the AMP standards and keeps a cached copy of their source code on Google’s web servers [347]. When a mobile user clicks on an AMP result in the SERPs, they automatically receive the cached copy from Google’s web server on its browser [271]. As a result, mobile users immediately receive the information they are looking for without any further network delay. Web pages that follow the AMP standards rank higher in mobile SERPs [22]. Although AMP technology is effective for both search engine rankings and web page performance, achieving the AMP standards is a demanding task [27,348].

Google locates and evaluates pages that adhere to AMP standards on a daily basis. Once Google identifies a page as AMP-compliant, it stores a cached copy of the page code on its servers. This caching enables the page to appear more quickly in search results. While AMP pages tend to rank higher in searches, visitors will view the content from Google’s web server, and their visit won’t be counted in the actual website traffic. Although the primary concern of SEO is to increase website traffic, utilizing AMP technologies adds value to the business’s brand rather than directly boosting organic traffic to the server. **Figure 46** illustrates the differences in search results for pages rated as AMP and those that are not.



**Figure 46.** Search results without and with AMP

Each AMP page must adhere to a specific format to successfully pass Google Bots AMP testing [23]. Initially, the page should contain Meta Data with structured data, specifically in JSON-LD format. Navigation should be structured using the “amp-sidebar” component. Images should follow a specific format, as should the price of the product [349].

When creating an AMP page, a canonical link and an amp link, respectively, as shown in **Figure 47**, should be added to the AMP header and the non-AMP page header. Canonical and amp links essentially inform search engines, that besides the AMP page, there is also a non-AMP page. Without the canonical and amphtml link, the AMP page can not be discovered by search engines [24].

```

//AMP Page Source
<!doctype html>
<html ⚡="" lang="en">
<head>
  <meta charset="utf-8">
  <link rel="canonical" href="https://www.website.com/product/macbook-pro-mj1q2/">
</head>

//Website Page Source
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="utf-8">
  <link rel="amphtml" href="https://www.website.com/product/macbook-pro-mj1q2/amp">
</head>

```

Figure 47. Canonical and Amp html link on header

To verify if a generated page is an AMP Page, Google has developed a tool [349]. The Washington Post emphasizes the significance of AMP pages, stating, “If our site takes a long time to load, it doesn't matter how great our journalism is, some people will leave the page before they see what's there”.

Accelerated Mobile Pages (AMP) are covered in more detail in **Chapter 5**.

#### 4.4.6. Apache Deflate and Gzip

There instances where websites load extremely slowly, leading users to become discouraged and choose to leave. The loading time of a website (page load time) is increasingly crucial for search engines when assessing its optimization [277].

The process of loading a website starts when the user, through their browser, requests a web address – URL. The web server receives the request and sends back the requested page. If the page is too “heavy”, the web server may delay in sending it to the user's browser, resulting in a prolonged waiting time for the user. In such cases, users might opt to look for another page that provides the desired information more efficiently and quickly, contributing to the bounce rate.. The average recommended page size is 150 kilobytes [292].

Two advanced and reliable solutions that can be applied to speed up websites are Deflate in Apache Server or Gzip on any other web server software solution [350]. In both cases, the content is compressed on the web server and sent back to the user's browser in a compressed form. The browser then handles the decompression and displays the content to the user. With this technique, users receive the page up to 70% faster. Since the page load time is shorter, users are less likely to abandon the website due to slow loading, and search engines may reward the site with a higher ranking [292,351].

#### 4.4.7. Page Caching

Client-side Page Caching, in turn, enhances website loading speed. Instead of compressing the page on the web server and sending it to the browser, it retrieves the page from the web server and temporarily stores specific files in the user's browser cache. When the user revisits the page, the browser loads the stored files from the cache and only requests the rest from the web server. This results in a quicker loading time on subsequent visits.

Additionally, there are server-side caching approaches that generate a cached copy of the produced HTML in a temporary location until the page is modified. Users receive a cached copy of the page without waiting for the scripting language to retrieve data from the database. A great but not limited example of server-side caching is the TWIG template engine [352].

Client-side and Server-side caching approaches can work simultaneously without interfering with each other. Both contribute to a faster-loading website , leading to higher rankings [292,353].

#### 4.4.8. Minify Resources (HTML, CSS, and JavaScript)

Minification refers to the process of removing unnecessary data without affecting how the resource is processed by the browser, for example, removing code comments, formatting, shorter variables, and function names [271]. In addition to optimizing images, as described in **Section 4.2.5**, all other files on a website should be as lightweight as possible to enhance page load time and user

experience (UX) [20]. Javascript (.js), Cascading Style Sheets (.css), and HyperText Markup Language (.html) files can be minified in order to improve a website's load time and SEO results. Performance evaluation studies have shown that minification can supply up to a 16% decrease in load time and up to 70% smaller files' total size [308]. Beyond enhancing website speed, minification can also contribute to the security of a website. During the minification process, variable and function names in JS code are replaced by short and meaningless names, thus, making the code extremely unreadable and human-unfriendly [351]. **Figure 48** provides an example of a file from the jQuery library before and after minification.

<b>Uncompressed jQuery 3.3.1</b> <i>Filesize: 272kb</i>	<b>Compressed jQuery 3.3.1</b> <i>Filesize: 87kb</i>
<pre> /*!  * jQuery JavaScript Library v3.3.1  * https://jquery.com/  *  * Includes Sizzle.js  * https://sizzlejs.com/  *  * Copyright JS Foundation and other contributors  * Released under the MIT license  * https://jquery.org/license  *  * Date: 2018-01-20T17:24Z  */ (function( global, factory ) {     "use strict";     if ( typeof module === "object" &amp;&amp; typeof module.exports === "object" ) {         // For CommonJS and CommonJS-like environments where a proper `window` // is present, execute the factory and get jQuery. // For environments that do not have a `window` with a `document` // (such as Node.js), expose a </pre>	<pre> /*! jQuery v3.3.1   (c) JS Foundation and other contributors   jquery.org/license */ !function(e,t){"use strict";"object"===typeof module&amp;&amp;"object"===typeof module.exports? module.exports=e.document?t(e,!0):function(e) {if(!e.document)throw new Error("jQuery requires a window with a document");return t(e)}:t(e)} ("undefined"!==typeof window? window:this,function(e,t){"use strict";var n= [],r=e.document,i=Object.getPrototypeOf,o=n.slice ,a=n.concat,s=n.push,u=n.indexOf,l= {},c=l.toString,f=l.hasOwnProperty,p=f.toString,d= p.call(Object),h={},g=function e(t) {return"function"===typeof t&amp;&amp;"number"!==typeof t.nodeType},y=function e(t){return null!=t&amp;&amp;t===t.window},v= {type:!0,src:!0,noModule:!0};function m(e,t,n) {var i,o= (t=t  r).createElement("script");if(o.text=e,n)fo r(i in v)n[i]&amp;&amp; (o[i]=n[i]);t.head.appendChild(o).parentNode.remov eChild(o);function x(e){return null==e? e+"":"object"===typeof e  "function"===typeof e? l[c.call(e)]  "object":typeof e}var b="3.3.1",w=function(e,t){return new w.fn.init(e,t),T=RegExp("\\s\\uFEFF\\xA0 +  \\s\\uFEFF\\xA0 +\$/g;w.fn=w.prototype= </pre>

**Figure 48.** Filesize Uncompressed & Compressed jQuery

#### 4.4.9. Open Graph Protocol (OGP)

Open Graph (OGP) is also a form of structured data developed by Facebook to enable the integration of external content into the social networking platform [346]. The OGP enables any web page to become a rich object in a social graph. For instance, OGP is used on Facebook to allow any web page to have the same functionality as any other object on Facebook [354]. Once a website's URL is shared on social media, social media bots visit the website trying to find three key elements of the web page, such as title, image, and short description.

With the integration of OGP, webmasters pre-markup these elements and make it easier for social media bots to display the result they are looking for. The use of OGP, like any other structured data format, provides richer results to search engines and social media end-users and is a crucial factor for a user to decide whether to click on and visit the corresponding website or not [18].

OG are essentially meta tags, each one containing properties and contents, placed in the header of the page, preferably after its title. The most well-known properties are `og:title`, `og:description`, `og:type`, `og:url`, `og:image` [282]. OG is also considered structured data, like Microdata; however they can be applied either simultaneously or independently, too. The latter is depicted in **Figure 49**.

```

<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="utf-8">
<title>Macbook Pro MJLQ2</title>

<!-- OG Tags -->
<meta property="og:url" content="https://www.website.com/product/macbook-pro-mjlq2/" />
<meta property="og:type" content="website" />
<meta property="og:title" content="Macbook Pro MJLQ2" />
<meta property="og:description" content="The Apple MJLQ2 MacBook Pro with Retina Display
has a compact and suave design that accommodates immense power." />
<meta property="og:image" content="https://www.website.com/img/macbook-pro-mjlq2.jpg" />
</head>

```

Figure 49. The Opengraph Protocol Example

In terms of SEO, OG meta tags play a crucial role. Web crawlers routinely visit millions of websites, and if they observe that a site consistently generates new content, they tend to visit it more frequently. Since social media users regularly create fresh content, web crawlers spend significant time searching for new content on these platforms. Therefore, when website content is republished on social media using OG, there is a higher likelihood that search engines will index this content. Moreover, social media generates substantial traffic, leading republished content to garner more clicks and websites to attract additional visitors (Figure 50).

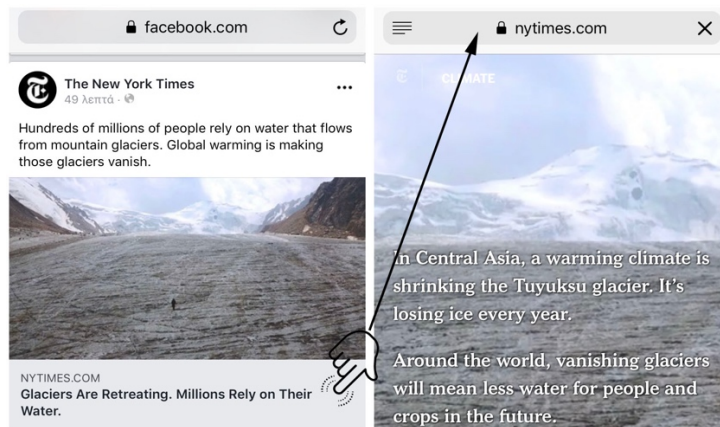


Figure 50. Facebook page post using OG, earn visitor to website

#### 4.4.10. Content Delivery Network (CDN) Links

A content delivery network (CDN) is a globally distributed network of proxy servers, known as Points of Presence (POPs), strategically placed around the world, so that end users can access the Internet content with low latency and high Quality of Experience (QoE) [312]. Its primary purpose is to store and create cached copies of assets (HTML, JS, CSS, images, etc.) of websites. When a user visits a website, these assets are served by the POP that is closest to the user, resulting in faster service to the end-user, due to reduced propagation delay, from the shorter physical distance. Many content providers offer CDN services, providing options such as monthly payment packages or pay-as-you-go packages, as exemplified by Google's Cloud CDN.

While many companies desire to adopt technologies to improve the page load speed of their website, some may not have the resources to invest in CDNs with sufficient bandwidth to cover their website traffic. A cost-effective solution is offered by Cloudflare, which provides free CDNs with limited bandwidth [355].

Alternatively, websites can opt for partially CDN adoption by using CDN URLs to load libraries such as jQuery CDN and Bootstrap CDN. For instance, a website could utilize the CDN provided by the "jquery.com" website to load the jQuery, saving both space and traffic on its web server. Visitors would then receive the jQuery library from CDN Url, i.e., from the closest POP Server to them. Figure 51 illustrates a partial adoption of the CDN, showcasing examples of Bootstrap, jQuery, and Font Awesome CSS libraries.

```
<!-- Bootstrap CSS CDN -->
<link href="https://stackpath.bootstrapcdn.com/bootstrap/4.1.3/css/bootstrap.min.css">
<!-- jQuery JS CDN -->
<script src="https://code.jquery.com/jquery-3.3.1.min.js"></script>
<!-- Font Awesome CSS CDN -->
<link href="https://use.fontawesome.com/releases/v5.5.0/css/all.css">
```

Figure 51. Partial adoption of CDN (Bootstrap, jQuery and Font Awesome libraries)

PageSpeed is an equally important ranking factor for a website [356]. In its 2010 announcements, Google informed webmasters through its Webmaster Central Blog that “Speeding up websites is of vital importance – not just to site owners, but to all Internet users. Faster sites create happy users and we’ve seen in our internal studies that when a site responds slowly, visitors spend less time there” [357].

#### 4.5. Black Hat SEO Techniques

White hat SEO is a term used to refer to SEO strategies and techniques that operate within the rules and expectations of search engines and search users [358].

The impact of white hat SEO requires considerable effort from website maintainers and manifests gradually [359]. To achieve higher rankings in SERPs in a shorter period, black hat SEO techniques have been devised [284]. Black hat SEO techniques violate the rules and promote spam websites by deceiving users and search engines [285]. While this study is dedicated to revealing various White Hat SEO techniques as a review, a brief overview of the most common black hat SEO techniques are also presented below.

- Content Automation – Content Spinning: Websites have to create valuable content to survive. This content is evaluated by the search engines, and the web page is indexed to the SERPs for specific keyword queries. Content Automation is a procedure where multiple articles are combined into one, while an automation tool attempts to spin the position of words to achieve a unique outcome [285].
- Cloaking: Cloaking is a black hat SEO technique used by Webmasters to trap visitors and search engines by presenting a different copy of content to users than to the search engines [286]. In this way, they manage to achieve positions in searches on specific high-traffic keywords, while irrelevant content is presented to their users [286].
- Keyword Stuffing: Keyword Stuffing is also a common black hat SEO technique in which keywords are placed multiple times within the web page in an attempt to mislead search engines about the content [360].
- Private Blog Networks (PBN): PBN are groups of high DA websites used exclusively for link building. Each PBN website links, by using backlinks, to the site they want to boost in the search results [361].

#### 4.6. Paid Traffic

Many websites cannot afford to wait a long time to attract organic visitors to their content. Acquiring organic visitors is a gradual process that requires time, thoughtful design, high-quality content, and effective SEO strategies. To address this challenge, a common solution is to purchase paid traffic, which essentially drives visitors to the website for a fee. Websites with significant traffic, such as social networking pages, can send thousands of visitors to a website in just a few hours, depending on the budget allocated for the paid traffic [288].

Search engines also provide opportunities for paid traffic. Webmasters can specify the number of users want to visit their website by choosing keywords and engaging in pay-per-click (PPC) advertising. Knowing that a website has paid to appear first in the searches for a particular keyword, search engines place it at the top of the search results. Following this process, the website increases its

traffic through paid ads. However, to maximize results, targeted campaigns must be created, consisting of several high-volume keywords.

While paid traffic is not a SEO methodology, it may be necessary for websites that lack the time-luxury to invest in SEO techniques. Combining paid traffic and SEO strategies can certainly yield even better results.

## ***4.7. A Machine Learning Python-Based Search Engine Optimization Audit Software***

### **4.7.1. Section Summary and Significance of the Study**

In the field of website traffic optimization, a need has emerged to adhere to search engine guidelines in order to attain a higher position in search results. Without a doubt, for more than two decades, the research productivity to highlight a plurality of SEO techniques and their importance in obtaining improved search rankings has been established in a clear manner. Nevertheless, there is a very limited amount of publications that recommend in an explicit manner which techniques can be employed by a web admin and in what order to maximize their website's SEO outcomes.

Under these circumstances, the necessity to optimize websites following the search engine guidelines has arisen, with the ultimate aim of attaining a higher position in search results, and hence greater possibilities to receive clicks and organic search traffic. Prior research studies have successfully identified the available SEO techniques and their significance in achieving better positions in search results. However, none of these publications explicitly recommend which SEO techniques website owners should utilize and in what sequence to optimize their SEO outcomes.

Commercial SEO audit tools have been developed to address this gap by scanning the source code of websites to identify implemented SEO techniques and detect any shortcomings. These tools are offered in freemium versions that enable users to monitor a single web page and encourage them to purchase the SEO tool to extend scanning to additional websites. As a result, website owners are required to pay monthly to identify their competitors' strengths and weaknesses, regardless of whether they are in the early stages or have an established online presence. While the cost may be manageable for profit-generating websites, it can be overwhelming for new websites or non-profit organizations. Startups possess an even more pronounced imperative for engaging in SEO to swiftly capture visitors, thereby fostering the sustainability of their ventures [362].

The present study aims to develop an open-source Python-based SEO audit software tool that will be accessible to the general public without charge and perform functions comparable to those offered by commercial SEO audit tools at a cost. The overarching objective is to produce an open-source SEO tool that will provide users with recommendations on appropriate SEO techniques based on analyzing their competitors' websites to optimize their websites for SEO and achieve improved search rankings and traffic.

**Section 4.7.2** explains the process and the rationale behind the creation of the Python-based SEO tool, which analyzes the website's source code and competitor websites. Free APIs are used to extend the tool's functionality and ML to predict off-page SEO techniques and critical metrics.

**Section 4.7.3** encompasses the utilization of an SEO tool to examine the competitive landscape of an operational e-commerce website. The findings suggested the incorporation of additional SEO techniques into the website's source code to enhance its prominence within search engine rankings. The findings of the study were remarkably significant, as the website experienced a substantial average increase of 143 daily organic visitors following the implementation of the recommended enhancements derived from the SEO tool. The incremented organic traffic coming from the proposed SEO tool is noteworthy since the corresponding previous average was merely ten daily organic visitors obtained from search engine results before the integration of our enhancements.

The proposed tool is designed to cater to a wide range of audiences, including website owners, digital marketers, and SEO professionals who are seeking to optimize their websites for improved search rankings and increased traffic. By utilizing our open-source Python-based SEO audit software tool, users can gain valuable recommendations on appropriate SEO techniques tailored to their specific needs. An important aspect worth highlighting is that even individuals lacking prior knowledge of SEO principles can benefit from the tool's comprehensive instructions, empowering them to independently implement the suggested SEO modifications on their websites.

One of the main advantages of the software developed for the purposes of this study, compared to commercial tools, is its initial open-source nature, which makes it accessible to individuals with limited budgets or non-profit organizations. This particular software provides an all-in-one solution by offering recommendations for both on-page and off-page SEO techniques that a website should follow to outperform its competitors [363]. In contrast, commercial tools specialize either in on-page or off-page SEO, forcing users to purchase multiple services.

#### **4.7.2. Materials and Methods**

This section presents the approaches and methodology followed to create the SEO Audit software. Furthermore, a comprehensive depiction is provided regarding the Python packages, external APIs, and machine-learning techniques utilized during the software development process. It is noteworthy to mention that the software developed for the purposes of this study is openly accessible as an open-source repository on GitHub [364].

The methodology followed for the creation of the software, model training, and the utilization of the software in a specific case study is described in the following 15 steps:

1. Identification of on-page and off-page SEO techniques;
2. Development of functions that detect SEO techniques in the source code of the webpage for each SEO technique;
3. Creation of classes containing SEO techniques;
4. Integration of the software with a third-party API for fetching SERPs;
5. Integration of the software with third-party APIs for gathering data related to speed, responsive design, and DA;
6. Generation of a dataset from live websites with measurements for DA, Linking Domains, and Backlinks;
7. Training of the Random Forest Regression model to predict Linking Domains and Backlinks based on DA;
8. Evaluation of the model's results;
9. Selection of a high-traffic relevant keyword;
10. Collection of ranking data for websites ranking on the first page of SERPs for the specific keyword;
11. Competitor analysis using the SEO audit tool;
12. Assessment of a live e-commerce site before implementing SEO techniques and keyword targeting;
13. Application of the SEO techniques suggested by the software audit tool to the target e-commerce site;
14. Collection of ranking data after 85 days of implementing SEO techniques on the targeted e-commerce site;
15. Evaluation of traffic results from Google Analytics for the e-commerce site.

#### **SEO Tool Functionality and APIs**

The current study presents an SEO tool that requires the provision of a live website's URL and a target keyword by the user, with the objective of attaining a first-page ranking for the specified keyword. To attain such a ranking, both on-page and off-page SEO techniques must be superior to or at least comparable with those employed by competitors. The SEO tool, therefore, adopts a

methodology that entails the detection of competing websites for a specific keyword, followed by a comprehensive competitor analysis of their on-page SEO techniques. The tool then employs free APIs and ML to identify competitors' off-page techniques. Once these data are obtained, the SEO tool scans the user's website to identify both on-page and off-page SEO techniques following a similar methodology. Equipped with knowledge regarding the strategies employed by competing entities, the SEO tool proffers SEO recommendations for the website at hand. Implementation of these suggested solutions has the potential to facilitate a first-page ranking and increase the target website's traffic.

The following sections present the methodology and rationale for implementing the SEO tool.

#### *Retrieve Search Engine Result Pages from Google (SERPs)*

The intention was to utilize a software program for scraping Google's search results. However, while developing the corresponding method, it was discovered that multiple requests from the same IP were detectable by Google's search engine, leading to the software being blocked after a few requests. To mitigate this challenge, proxies were integrated into the SEO tool, allowing for the rotation of IP addresses at regular intervals. While the free-proxy package [365] was functional, it primarily utilized free public proxies often already blocked by Google. The purchase of premium proxies was considered but conflicted with the study's principles as an open-source solution.

Consequently, the ZenSerp API [366] was employed, which retrieves Google search results in JSON format through an API request. Users can access up to 50 API requests per month without a subscription by creating an account. Upon receiving a target keyword from the user, the SEO tool utilizes the ZenSerp API to make a request, resulting in the first-page search results being returned in a JSON-structured format. These results are then processed and stored in a Python dictionary.

#### *SEO Techniques Methods and Python Packages*

The code employs several libraries to enable its functionalities:

- The `requests` library is utilized to make API requests, while the `JSON` library is used to convert JSON data into a Python dictionary.
- The `urllib.parse` library is used to parse the URL in the `get_robots(URL)` function.
- The `re` library, which supports regular expressions, is also used in the code.
- The `csv` library is used to save the dictionary to a CSV file.
- The `BeautifulSoup` library is also employed for web scraping purposes, allowing the code to extract data from HTML files.
- Finally, the `Mozscape` library obtains a website's DA.

These libraries enable the code to perform various tasks, from making API requests and parsing URLs to scraping websites and retrieving data on DA.

The SEO audit software described in this section comprehensively analyzes each website listed in the dictionary. To achieve this, the tool employs a web scraping technique that involves searching the website's source code for on-page SEO techniques. The tool utilizes 24 individual SEO metrics that are briefly described in **Sections 4.2, 4.3, and 4.4**, with each one designed to identify a specific on-page SEO technique within the code. A comprehensive list of these techniques and associated methods are provided in **Table 18**. The tool's ability to identify and analyze on-page SEO techniques plays a critical role in its effectiveness in proposing solutions for website optimization and ranking improvement.

**Table 18.** Methods and their corresponding SEO Techniques.

Method	SEO Technique	Description
<code>perform_seo_checks</code>	-	The method uses the <code>requests</code> method to obtain the website's source code and the <code>BeautifulSoup</code> library as <code>html.parser</code> to parse the code. Finally, it utilizes all the

		methods that detect SEO techniques by returning the results.
<b>get_organic_serps</b>	-	The method utilizes the ZenSerp API to retrieve a comprehensive list of websites that are listed on the first page of search results in response to a given keyword query.
<b>get_image_alt</b>	Image Alternative Attribute	The method is designed to identify and flag instances of missing alternative attributes for images from a provided list.
<b>get_links_title</b>	Link Title Attribute	The method involves the identification of missing title attributes within a list of links.
<b>get_h_text</b>	Heading 1 and 2 Tags	The method detects and counts the occurrences of H1 and H2 tags within the source code, followed by a search for the specified target keyword within these tags.
<b>get_title_text</b>	Title Tag	The method conducts a search for the presence of the title tag within the source code and verifies the existence of the target keyword in conjunction with the tag.
<b>get_meta_description</b>	Meta Description	The method conducts a search for the presence of the meta description within the source code and verifies the existence of the target keyword in conjunction with the tag.
<b>get_meta_opengraph</b>	Opegraph	The method conducts a search for the presence of the opengraph tag within the source code and verifies the existence of the target keyword in conjunction with the tag.
<b>get_meta_responsive</b>	Responsive Tag	The method conducts a search for the presence of the viewport within the source code and verifies the existence of the target keyword in conjunction with the tag.
<b>get_style_list</b>	Minified CSS	The method identifies the stylesheets present in the source code and examines whether they have been minified.
<b>get_script_list</b>	Minified JS	The method identifies the scripts present in the source code and examines whether they have been minified.
<b>get_sitemap</b>	Sitemap	The method identifies the xml sitemap present in the source code.
<b>get_rss</b>	RSS	The method identifies the RSS feed present in the source code.
<b>get_json_ld</b>	JSON-LD structured data	The method identifies the JSON-LD present in the source code.
<b>get_item_type_flag</b>	Microdata structured data	The method identifies the Microdata present in the source code.
<b>get_rdfa_flag</b>	RDFa structured data	The method identifies RDFa present in the source code.
<b>get_inline_css_flag</b>	In-line CSS	The method detects any in-line CSS code in source code.
<b>get_robots</b>	Robots.txt	The method verifies the presence of a Robots.txt file in the root path of the website.
<b>get_gzip</b>	GZip	The method performs an evaluation to detect the presence of gzip Content-Encoding in response headers.
<b>get_web_ssl</b>	SSL Certificates	The method performs an analysis to determine if the webpage has been provided with Secure Sockets Layer

		(SSL) certificates, which ensure that the connection between the user's browser and the website is encrypted and secure.
<b>seo_friendly_url</b>	SEO Friendly Url	The method examines whether the provided URL is optimized for SEO, conforming to the best practices and guidelines.
<b>get_speed</b>	Loading Time	The method employs the Lighthouse API to measure the web page's loading time.
<b>get_responsive_test</b>	Responsive Design	The method employs the mobileFriendlyTest API to ascertain whether a given webpage is responsive, i.e., capable of rendering suitably on different devices and screen sizes.
<b>get_da</b>	DA	The method employs the MOZ API to obtain the DA of a website.

### *External APIs*

To enhance the functionality of the SEO tool, four APIs were integrated into its code to perform a thorough analysis of the targeted website. These APIs include the PageSpeed Insights API, Mobile-Friendly Test Tool API, Mozscape API, and the Google SERP API. By utilizing these APIs, the SEO tool can provide valuable insights into different aspects of the website's performance, including its mobile usability, page speed, search engine ranking, and DA. Integrating these APIs helps the SEO tool perform a deep analysis of the website and identify areas for improvement in terms of on-page and off-page SEO techniques. Additionally, using these APIs allows the SEO tool to provide specific recommendations on addressing any issues found during the analysis. Overall, integrating these APIs dramatically enhances the functionality and effectiveness of the SEO tool, enabling website owners to optimize their website's performance and improve its visibility on Search Engine Result Pages (SERPs).

- Mobile-Friendly Test Tool API, developed by Google, is a web-based service verifying a URL for mobile-usability issues. Specifically, the API assesses the URL against responsive design techniques and identifies any problems that could impact users visiting the page on a mobile device. The assessment results are then presented to the user in the form of a list, allowing for targeted optimizations to improve the website's mobile-friendliness [338].
- PageSpeed Insights API, developed by Google, is a web-based tool designed to measure the performance of a given web page. The API provides users with a comprehensive analysis of the page's performance, including metrics related to page speed, accessibility, and SEO. The tool can identify potential performance issues and return suggestions on optimizing the page's performance. This allows website owners to make informed and specific decisions regarding speed optimization that can be made to improve user experience and overall page performance [366].
- Mozscape API, The Mozscape API, developed by MOZ, is a web-based service that provides accurate metrics related to a website's performance. Specifically, the API takes a website's URL as input and returns a range of metrics, including DA. DA is a proprietary metric developed by MOZ that measures the strength of a website's overall link profile. The metric is calculated using a complex algorithm that considers various factors, such as the quality and quantity of inbound links. It objectively assesses a website's authority relative to its competitors. The Mozscape API is a valuable tool for website owners and SEO professionals seeking insights into their website's performance and improving their overall search engine rankings [367].
- Google SERP API, developed by ZenSerp API, is a web-based service that allows users to efficiently and accurately scrape search results from Google. The API is designed to provide users with a seamless experience, offering features such as rotating IP addresses to prevent

detection and blocking by Google and returning search results in a JSON-structured format. The tool is a valuable asset for SEO professionals and website owners seeking to gain insights into their website's performance and improve their search engine rankings. The Google SERP API developed by ZenSerp API is an efficient and reliable tool for web scraping, offering accurate search results and facilitating the process of data analysis and SEO optimization [366].

### **Machine Learning**

Two additional factors that must be considered for ranking a website in searches, as mentioned in **Section 4.7.1**, are the number of backlinks and linking domains. Backlinks refer to the number of links from third-party pages that point to the target page, while linking domains refer to the number of unique domain names that point to the target page.

In the **External APIs** section, the Mozscape API was utilized to acquire the DA of both the user's webpage and its competitors. However, the DA is computed through intricate algorithms that consider the authority of each backlink directed toward the page. Gathering, collating, and storing links for each website incur significant costs for the providers, amounting to millions of dollars. Consequently, these companies offer limited information regarding the number of backlinks and linking domains only for individual searches rather than in an API format. While collecting this information through an API is feasible, it often involves a cost exceeding \$3000 per year and limitations on the Rows returned by the API.

Rather than subscribing to an external service—that could burden significantly the budget of an organization independently of its size—our approach involved developing two prediction models employing ML techniques. These models were designed to forecast the number of backlinks and linking domains based on the DA. To train these models, we manually collected data on the DA, the number of backlinks, and the number of linking domains for a sample of 150 live websites selected from the DMOz Open Directory Project (ODP). The data was compiled in a CSV file and utilized for model training. Subsequently, pre-trained files were generated from the trained models to enable the SEO tool to expedite predictions.

#### *Model Training*

**Figure 52** depicts the methodology employed for model training and the generation of pre-trained models employed by the SEO software. The implementation relies on the utilization of three distinct libraries.

1. Pandas is a widely used Python library that offers an extensive array of functions for manipulating and analyzing data. Its functionalities include support for data frames and series, which enables structured data processing [368]. The software employs pandas to read the data from a CSV file, preprocess it, and create new data frames to store the independent and dependent variables.
2. Scikit-learn is a machine-learning Python library that offers a diverse set of tools for data analysis and modeling [369]. In this software, scikit-learn is employed to train and evaluate the Random Forest Regression models.
  - a. RandomForestRegressor is a class implemented in the scikit-learn library that embodies the Random Forest Regression algorithm [369]. This class is utilized in the code to train Random Forest Regression models, which are utilized to make predictions of the dependent variables, namely Backlinks and LinkingDomains, based on the independent variable, DA. The RandomForestRegressor technique operates as an ensemble approach, seamlessly blending numerous decision trees to forge a sturdy and precise model. Its versatility spans both regression and classification tasks, rendering it a fitting selection for prognosticating numerical metrics such as backlinks and linking domains. Also, they are less prone to overfitting compared to individual decision trees. They create multiple trees and aggregate their predictions, reducing the risk of learning noise in the data. This can

lead to more reliable and stable predictions, which is crucial when dealing with real-world data. Finally, by training multiple trees and combining their predictions, random forests tend to be less sensitive to fluctuations in the dataset, resulting in a more consistent performance across different subsets of data.

- b. `train_test_split` is a function incorporated in the `scikit-learn` library, which is employed to partition the dataset into training and testing sets [369]. This function randomly splits the data into two separate subsets, where one is used for training the ML model, and at the same time, the other is utilized for testing its performance.
  - c. The `mean_squared_error` function in `scikit-learn` is a mathematical function that calculates the mean squared error (MSE) between the actual and predicted values of the dependent variable. In the context of the presented software, this function is used to evaluate the performance of the trained Random Forest Regression models on the testing data. It measures the average squared difference between the actual and predicted values, where a lower MSE indicates a better fit of the model to the data [369].
3. `Joblib` is a Python library that provides tools for the efficient serialization and deserialization of Python objects [370]. In this software, it is used to store the trained models on disk and retrieve them later to make predictions on new data. By storing the models as files, the trained models can be shared and used in other applications without retraining. This also enables efficient storage and retrieval of models, which can be especially useful for larger models requiring significant computational resources.

During the model training phase, the first step involves loading a CSV file named `Dataset.csv` into a pandas DataFrame object `data`. Subsequently, the independent variable `DA` and dependent variables `Backlinks` and `LinkingDomains` are extracted from data and stored in separate pandas series labeled as `X`, `y1`, and `y2`, respectively.

Next, data is split into training and testing sets utilizing the `train_test_split` function of `scikit-learn`. Two Random Forest Regression models are then trained using the training data, one to predict `Backlinks` (`rf1`) and the other to predict `LinkingDomains` (`rf2`). Both models are trained using 100 trees and a random state of 42.

The trained models' performance is then assessed on the testing data, utilizing the MSE metric, which measures the average squared difference between the predicted and actual values of the dependent variable. The resulting MSE values are stored in `mse1` and `mse2` variables and printed to the console.

Finally, the pre-trained models are saved to disk using the `joblib.dump()` function. The function generates two separate files, named `da-to-backlinks.joblib` and `da-to-linking-domains.joblib`, which will be utilized in production mode for making predictions.

The `ipynb` files for model training, as well as the CSV and `joblib` pre-trained files, are available on GitHub [364].

1. To assess the model's outcomes for the specific task, we conducted two distinct regression analyses. The results of the regression analysis reveal valuable insights into the relationship between the `DA` and `LinkingDomains` variables. The p-values associated with the coefficients provide critical information about the statistical significance of each variable in the model. The p-value for the variable `DA` is exceptionally low at approximately  $8.85 \times 10^{-98}$  indicating an extremely high level of statistical significance. This suggests a strong relationship between the `DA` variable and the predicted `LinkingDomains` values. The R-squared value of 0.93 further reinforces the model's effectiveness in explaining the variation in the dependent variable. With an R-squared value close to 1, it can be inferred that around 93% of the variation in `LinkingDomains` is accounted for by the linear relationship with `DA`. This robust R-squared value signifies that the regression model

provides a compelling fit to the data, underlining its predictive capability and potential for insights into the relationship between these variables.

2. The regression analysis conducted on the relationship between DA and Backlinks has also generated significant insights. The Ordinary Least Squares (OLS) regression model demonstrates a strong fit to the data, with an R-squared value of 0.630, implying that approximately 63% of the variability in Backlinks can be explained by changes in DA. The F-statistic of 282.7 with a corresponding p-value of  $1.10 \times 10^{-37}$  highlights the overall significance of the model, suggesting that the model as a whole is able to predict the Backlinks effectively.

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import joblib

# Load the data from CSV file
data = pd.read_csv('Dataset.csv')

# Preprocess the data
X = data[['DA']]
y1 = data['Backlinks']
y2 = data['LinkingDomains']

# Split the data into training and testing sets
X_train, X_test, y1_train, y1_test, y2_train, y2_test = train_test_split(X, y1, y2, test_size=0.2, random_state=42)

# Train a Random Forest model for Backlinks
rf1 = RandomForestRegressor(n_estimators=100, random_state=42)
rf1.fit(X_train, y1_train)

# Train a Random Forest model for LinkingDomains
rf2 = RandomForestRegressor(n_estimators=100, random_state=42)
rf2.fit(X_train, y2_train)

# Evaluate the models on the testing set
y1_pred = rf1.predict(X_test)
y2_pred = rf2.predict(X_test)
mse1 = mean_squared_error(y1_test, y1_pred)
mse2 = mean_squared_error(y2_test, y2_pred)
print(f'MSE for Backlinks: {mse1:.2f}')
print(f'MSE for LinkingDomains: {mse2:.2f}')

# Save the models to files
joblib.dump(rf1, 'da-to-backlinks.joblib')
joblib.dump(rf2, 'da-to-linking-domains.joblib')
```

Figure 52. Train Model Phase: DA to Backlinks and Linking Domain.

The results from the regression analysis are presented in scatterplots in Figure 53.

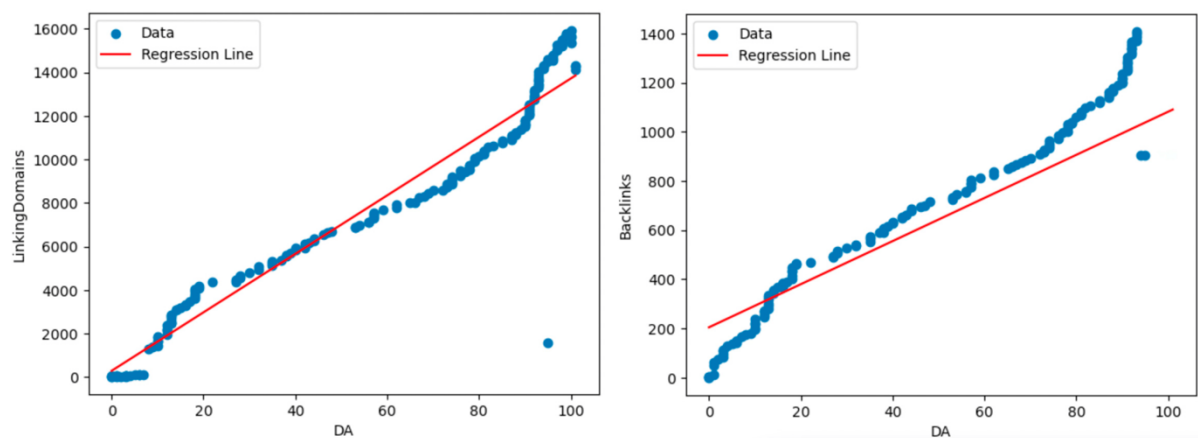


Figure 53. Scatterplots for LinkingDomains and Backlinks.

### Predictions Based on DA

In the Prediction Phase, a Python class called `Predict` is defined (Figure 54), consisting of two methods that take a value of `DA` as input and use pre-trained ML models to make predictions. The pre-trained models are loaded from disk using the `joblib.load` function and stored in the variables `rf1` and `rf2`. The `predict_backlinks` and `predict_linking_domains` methods utilize the `rf1` and `rf2` models, respectively, to predict the values for `Backlinks` and `LinkingDomains` for the new data point. The resulting predicted values for `Backlinks` and `LinkingDomains` are returned to the `SEOTechniques` file for further use.

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import joblib

# Load the models from files
rf1 = joblib.load('da-to-backlinks.joblib')
rf2 = joblib.load('da-to-linking-domains.joblib')

# Make a prediction for a new data point
new_data = pd.DataFrame({'DA': [50]})
backlinks_pred = rf1.predict(new_data)[0]
linkingdomains_pred = rf2.predict(new_data)[0]
print(f'Predicted Backlinks: {backlinks_pred:.0f}')
print(f'Predicted LinkingDomains: {linkingdomains_pred:.0f}')
```

Figure 54. Make Predictions Phase: DA to Backlinks and Linking Domain.

### 4.7.3. Results

Section 4.2 and 4.3 provides a concise and informative analysis of the on-page and off-page SEO techniques implemented by the tool, officially sanctioned by the Webmaster guidelines of the Google Search Central Documentation [271]. Additionally, Section 4.7.2 presents a comprehensive explanation regarding the development of the Python-based SEO tool, emphasizing the underlying processes. The tool’s functionality encompasses an in-depth SEO analysis of the website’s source code and competitor websites for a given keyword. The integration of freely available APIs was employed to expand its capabilities, while the incorporation of ML techniques facilitated the prediction of off-page SEO techniques and essential metrics.

In this section, the study undertook a comprehensive analysis of the competitive environment of an online e-commerce website using the SEO software. The principal aim was to propose additional SEO techniques to be integrated into the website’s source code, thereby augmenting its visibility and positioning within search engine rankings for a given keyword.

The software initially requested the URL of the target website and the target keyword. Utilizing the ZenSerp API, it retrieved the organic search results from Google using the target keyword as the search term. For each of the eight competitive websites (Table 19 C1 to C8), it identified the on-page SEO techniques and technologies they employed. Simultaneously, by making API requests to the Mobile-Friendly Test Tool, PageSpeed Insights, and Mozscape APIs, it extracted data related to the responsive SEO technique, web page speed, and DA for each website. Additionally, leveraging the pre-trained models and utilizing the DA data, predictions were made for each website’s number of Backlinks and Linking Domains.

Table 19. SEO Analysis for Competitors’ and User’s Websites (report-seo\_competitors\_table.csv).

SEO Techniques/Page	C1	C2	C3	C4	C5	C6	C7	C8	U
<b>position</b>	1	2	3	4	5	6	7	8	0
<b>amp</b>									
<b>images_alt</b>	✓			✓	✓		✓		✓
<b>links_title</b>	✓						✓		

heading1									
heading2									
title									
meta_description									
opengraph			✓			✓		✓	
style	✓						✓		
sitemap									
rss									
script	✓						✓		
json_ld			✓	✓				✓	
inline_css	✓						✓		
microdata					✓				
rdfa									
robots		✓		✓	✓	✓		✓	
gzip		✓	✓			✓		✓	
web_ssl	✓		✓	✓	✓	✓	✓	✓	✓
seo_friendly_url		✓	✓	✓	✓	✓		✓	✓
speed	2.9 s	1.4 s	1.5 s	0.4 s	3.8 s	1.7 s	1.9 s	1.3 s	2.8 s
responsive	✓	✓	✓	✓	✓	✓	✓	✓	✓
da	10	10	20	16	19	5	18	23	4
backlinks	1191	1191	17,324	2908	17,324	6	3261	6054	4
linking_domains	84	84	211	119	211	4	214	183	3

<sup>1</sup>C1: pasxalineslampades.com; C2: toolittle.gr; C3: armoniecandles.com; C4: tsago.gr; C5: e-gerakis.gr; C6: dekor.gr; C7: nuovavita.gr; C8: kerl.gr; U: messiniancandles.gr.

A similar procedure was employed for the URL of the provided e-commerce website (Table 19). Subsequent to the successful execution of the script, three CSV files were generated, each corresponding to Table 19, Table 20, Table 21 for presentation and analysis.

Table 20. SEO recommendations based on Competition (report-suggestions\_based\_on\_competition.csv).

Missing SEO Technique	Comments	SEO Rules
gzip	It is suggested to use the [gzip] SEO technique, which the majority of the competition have also applied.	<ul style="list-style-type: none"> <li>- Google Developers—Optimize Encoding and Transfer Size of Text-Based Assets: <a href="https://developers.google.com/web/fundamentals/performance/optimizing-content-efficiency/optimize-encoding-and-transfer">https://developers.google.com/web/fundamentals/performance/optimizing-content-efficiency/optimize-encoding-and-transfer</a></li> <li>- Mozilla Developer Network—HTTP compression: <a href="https://developer.mozilla.org/en-US/docs/Web/HTTP/Compression">https://developer.mozilla.org/en-US/docs/Web/HTTP/Compression</a></li> <li>- GTmetrix—What is Gzip Compression?: <a href="https://gtmetrix.com/enable-gzip-compression.html">https://gtmetrix.com/enable-gzip-compression.html</a></li> <li>- BetterExplained—The Importance Of Gzipping Your Website: <a href="https://betterexplained.com/articles/how-to-optimize-your-site-with-gzip-compression/">https://betterexplained.com/articles/how-to-optimize-your-site-with-gzip-compression/</a></li> </ul>

**Table 21.** SEO recommendations based on SEO tool (report-seo\_tool\_suggestions.csv).

Missing SEO Technique	Comments	Data	SEO Rules
links_title	The links on the WebPage are missing title tags. It is recommended to add title tags to all links for better user experience and SEO optimization.	{'/service/xondriki-polisi': "", 'product/bazo-130-mL-451': "", 'product/bazaki-me-fello-524': "", 'product/bazaki-me-fello': "", 'product/bazo-130mL-396': "", 'product/lampada-gamou-f12cm-20-cm': "", 'product/pasxalini-lampada-strogguli-krakele-mob': "", 'product/pasxalini-lampada-strogguli-krakele-prasini': "", 'product/lampada-gamou-tetragoni-masif-12cm12cm-22cm': "", 'product/parafinelaio': "", 'product/pasxalini-lampada-plake-xusti-sapio-milo': "", 'product/ekklisiastiko-keri-n1': "", 'product/kormos-apo-keri-f9cm95cm-455': "", 'product/parafini': "", 'product/keri-citronella-n1': "", 'product/premium-futika-keria-me-kapaki': "", 'product/potiri-vintage': "", 'product/potiri-stroggulo': "", 'product/kormos-apo-keri-9cm-9-cm-15cm': "", 'product/bazaki-me-fello-323': "", 'product/antipagotiko-keri': "", 'product/pasxalini-lampada-strogguli-mob': "", 'product/pasxalini-lampada-plake-krakele-prasini': "", 'product/lampada-baptisis-roz': "", 'product/akatergasto-prosanamma': "", 'product/pasxalini-lampada-strogguli-xusti-galazia': "", 'product/premium-futika-keria-me-kapaki-418': "", 'product/pasxalini-lampada-plake-xusti-galazia': "", 'product/pasxalini-lampada-plake-galazia': "", 'product/lampada-baptisis-galazia': "", 'product/set-kormon-apo-keri-339': "", 'product/ekklisiastiko-keri-n1-	<ul style="list-style-type: none"> <li>- Example: &lt;a href="..." title="..."&gt;</li> <li>- According to Google's SEO Starter Guide, adding descriptive title tags to your links can help both users and search engines understand the content of your pages. Google also considers title tags as a ranking factor for search results.</li> <li>- The WCAG also recommend using descriptive text for links, including title attributes, to ensure accessibility for people with disabilities.</li> <li>- Moz, a leading SEO software provider, also recommends using descriptive title tags for links as a best practice for SEO optimization.</li> </ul>

		500gr': '', 'product/potiri-funky-xalkino': '', 'product/potiri-koukounara': ''}	
--	--	--	--

**Table 19** showcases the SEO techniques along the y-axis alongside the respective ranking positions of the competitive websites within the organic search results for the specified keyword. The x-axis encompasses the eight competitive websites (C1-C8) and the URL of the e-commerce website (U). The presence of a check symbol signifies the successful implementation of particular SEO techniques on each respective webpage. Conversely, in instances where a web page has not incorporated the aforementioned SEO technique, the corresponding field remains devoid of any information. Consequently, **Table 19** facilitates the assessment of applied SEO techniques by the competition and the user's own website, providing valuable insights into their respective practices.

Considering the widespread adoption of a particular SEO technique by the majority of the competition, it is advisable for the user's website to incorporate the same, provided it has not already accomplished this. **Table 20** presents recommendations generated by the software pertaining to SEO techniques predominantly followed by the competition, which the user's website would benefit from implementing. In cases where a deficiency in the application of an SEO technique is identified, the tool offers comments and outlines SEO rules, accompanied by web references, to guide the user in effectively applying the recommended SEO technique.

In recognition of the fact that the techniques employed by the competition may not necessarily yield optimal outcomes, the software advises the user to implement any SEO techniques that are absent from their website, irrespective of whether or not they are followed by the competition (**Table 21**). By doing so, the user can enhance their website's overall SEO performance and potentially surpass the outcomes achieved by competitors. Due to the comprehensive nature of the suggestions generated by the SEO tool for the website, a concise representation is chosen, wherein only one recommendation out of the twelve proposed by the SEO tool is presented.

Based on the findings derived from the utilization of the SEO tool, it was observed that the website under scrutiny faced a substantial level of competition, as indicated by the presence of e-commerce entities with domain authorities exceeding 10. Concurrently, the analyses conducted unveiled that the predominant competitors enjoyed considerable support from SEO groups that had implemented advanced SEO techniques, such as `open_graph_protocols` or `json_ld` structured data. These sophisticated strategies contributed significantly to their enhanced SEO efforts.

Live e-commerce, as a new e-commerce venture, struggled to gain more than 10 organic visitors per day due to a lack of targeting specific high-traffic keywords. For this particular case study, the keyword "πασχαλινές λαμπάδες" (Easter candles) was identified, which was characterized as high traffic based on Google Ads measurements for the given period. The e-commerce site did not even appear in the top 10 pages of organic search results for this specific keyword, indicating a lack of keyword targeting as well as the implementation of on-page and off-page SEO techniques.

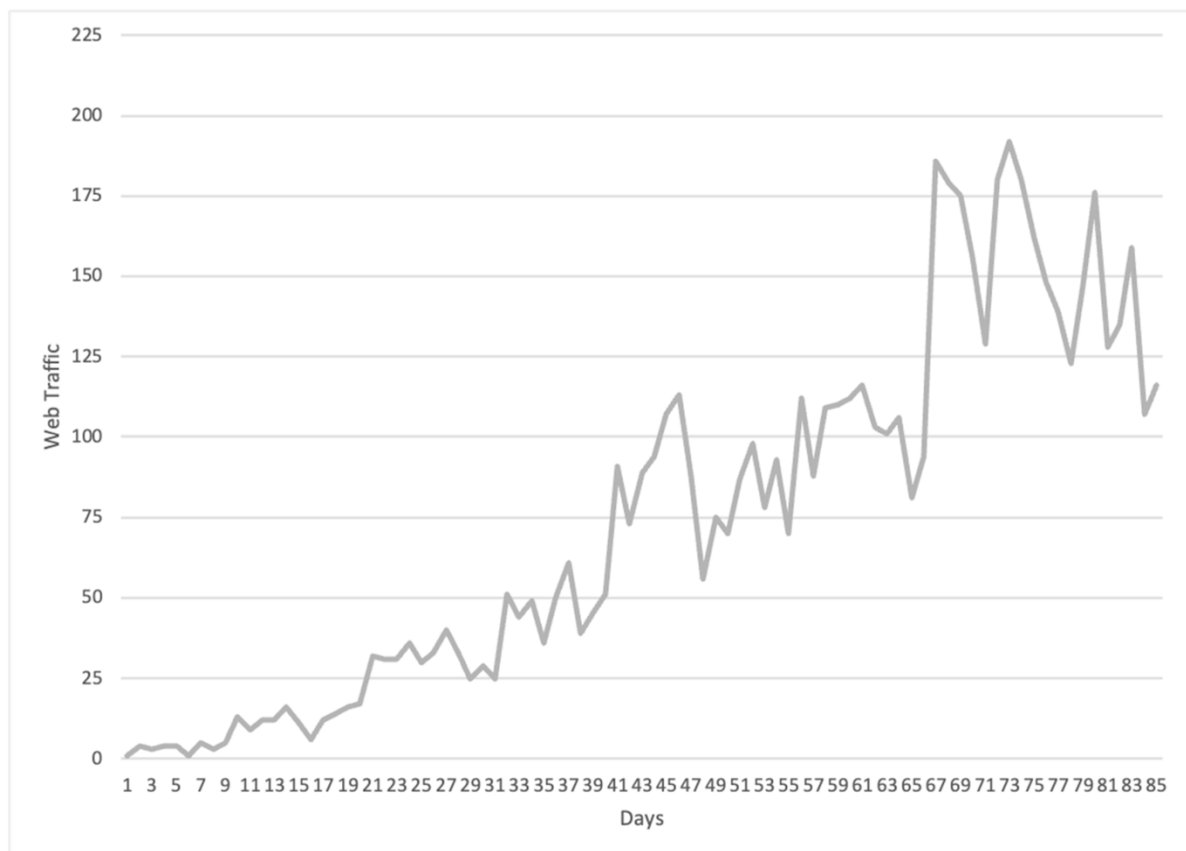
As stated in the introduction section, the cost for a web administrator to hire an SEO specialist is quite significant for a new e-commerce venture. This is further compounded by the supplementary costs requisite for the SEO specialist's utilization of premium SEO software, which, in turn, furnishes a repository of critical insights and data to more effectively propel the e-commerce enterprise. In contradistinction, the present software embodies an open-source, comprehensive solution capable of assisting each web administrator comprehensively in confronting competitive pressures. It encompasses both on-page and off-page SEO competitor analysis, alongside an SEO audit replete with recommendations tailored to enhance the performance of their respective websites and e-commerce platforms.

By adhering to the on-page and off-page SEO recommendations outlined by the SEO tool and incorporating a total of 17,324 backlinks originating from 211 distinct referring/linking domains, the e-commerce website achieved a notable placement on the first page of search results for the targeted keyword. Specifically, the website secured the fifth position in the organic search results. Web pages

that rank higher in search results, particularly on the first page of search listings, increase their chances of being clicked by searchers.

The presented results are focused on targeting a single keyword and could be multiplied if multiple keywords were targeted simultaneously.

To visualize the website's traffic patterns, **Figure 55** presents the traffic data recorded by Google Analytics, spanning from 26 February 2023 to 20 May 2023. The x-axis of **Figure 55** represents the duration, in days, required for the implemented SEO modifications to manifest noticeable effects on website traffic. Conversely, the y-axis denotes the corresponding levels of website traffic. Day 0 signifies the initial day following the implementation of all suggested SEO techniques recommended by the software, at which point the user anticipates the detection of these changes by the Google search engine and the subsequent repositioning of the web page to a more suitable/higher ranking. Remarkably, within a span of 85 days, subsequent to incorporating the suggested modifications proposed by the SEO tool, a substantial increase in traffic occurred, with an average of 143 additional visitors per day. Out of the total 6,270 visitors, the breakdown is as follows: 4,518 originated from organic search, 1,431 from direct traffic, 316 from paid traffic, and five from referral traffic. The attainment of first-page search rankings, along with the organic traffic increase constitutes a solid steppingstone to increase the online sales. A fact that we will continue to examine in our future efforts.



**Figure 55.** Google Analytics. Traffic trend from 26 February 2023 to 20 May 2023.

#### 4.7.4. Discussion

##### Importance of Accessibility and Affordability

The introduction of this open-source SEO audit software represents a transformative step towards democratizing access to advanced SEO analysis tools. By championing the values of accessibility and affordability, the software addresses a pressing industry need, empowering a diverse range of users. The traditional landscape, dominated by costly commercial solutions, has hindered

small businesses, non-profits, and individuals from harnessing the power of sophisticated SEO insights. This software shatters these barriers, providing a level playing field where data-driven optimizations are within reach for all. As a result, it not only empowers underrepresented entities to compete effectively in the digital realm but also fosters collaboration and innovation across the broader online community. Through this groundbreaking approach, the software marks a significant stride towards inclusivity and shared success in the dynamic world of digital marketing.

By dismantling the exclusivity that has long characterized the SEO tool landscape, the open-source software heralds a new era of empowerment and growth. Its emphasis on accessibility ensures that small businesses can navigate the digital landscape with newfound confidence, while non-profits can enhance their online impact without compromising their missions. Even individual website owners and entrepreneurs stand to benefit, amplifying their online presence through strategic optimizations previously reserved for larger enterprises. In this transformative shift, the software not only levels the playing field but also fuels a culture of collaboration and innovation. By embracing the principles of accessibility and affordability, the software champions the democratization of SEO analysis, heralding a brighter and more inclusive future for digital marketers of all backgrounds.

### **Integration of Traditional Programming and Machine Learning**

The software's evolution embodies an exceptional blend of traditional programming methods and state-of-the-art ML algorithms, a groundbreaking fusion that imparts it with a unique advantage in providing holistic on-page and off-page SEO strategies, unlike commercial tools that typically excel in either on-page or off-page SEO, but not both. This hybrid integration exemplifies a strategic leap forward, combining the proven principles of traditional coding with the adaptability and intelligence of ML.

By incorporating traditional programming, the software maintains a solid foundation rooted in established SEO principles. It executes meticulous analyses of on-page factors such as meta tags, keywords, structured data, etc. This approach caters to the essential groundwork of SEO, offering users a comprehensive evaluation of their website's foundational components.

On the other hand, the infusion of ML algorithms elevates the software's capabilities to a new realm of sophistication. ML's prowess in pattern recognition, data processing, and predictive modeling amplifies the software's analytical prowess. This integration allows the software to delve deeper into the complexities of off-page SEO, unraveling the intricacies of backlink profiles, and linking domains. The hybrid approach's adaptability empowers the software to evolve alongside the ever-changing landscape of search engine algorithms, making it a versatile tool that can remain effective even in the face of evolving SEO dynamics.

### **Micro-Level Insights and Tailored Guidance**

The software presents a dual-level approach that amalgamates micro-level insights and tailored guidance, ushering in a new era of precision-driven website optimization. At the micro-level, the software conducts meticulous analyses of individual SEO metrics, offering precise recommendations to address specific SEO issues. This empowers users to fine-tune their strategies with targeted actions, such as optimizing load speeds and every facet of their website's SEO performance.

Simultaneously, the software adopts a macro-level perspective that takes into account the bigger picture, aligning SEO efforts with a website's unique competition and goals. By identifying competition and SEO techniques employed, it creates an SEO strategy that is more suitable for the specific business sector.

This dual-level functionality not only diagnoses SEO challenges but also equips users with actionable roadmaps for success, transforming the software into a dynamic and personalized optimization companion that enhances search rankings, drives organic traffic, and elevates user engagement.

## Competitor Analysis and SEO Methodologies

In contrast to commercial tools, the software's unique competitive edge stems from its sophisticated competitor analysis capabilities, propelling its function beyond mere self-assessment. For instance, platforms like SeoSiteCheckup [371] and Seobility [372] focus exclusively on identifying and suggesting on-page SEO techniques applied to a web page, without considering the competitive landscape. In contrast, the Ahref Backlink Checker platform [373] solely concerns itself with identifying backlinks for a specific web page, without proposing, based on DA, the backlinks that a website might require to surpass its competition. This feature enables the software to dissect rivals' intricate SEO methodologies, offering users a deep understanding of their competitors' strategies. By extracting actionable insights from this analysis, the software equips users with strategic guidance to enhance their own SEO tactics. In a fiercely competitive e-commerce market, the software's competitor analysis unearthed strategic opportunities previously overlooked. By implementing the software's recommendations, the e-commerce platform witnessed a substantial boost in organic traffic. This success story underscores the software's potential to revolutionize SEO strategies, not only optimizing websites but also empowering users to outperform rivals. In essence, the open-source SEO audit software's ability to unravel competitor SEO methods and translate them into actionable advice sets it apart as a pivotal tool for strategic growth in the dynamic digital landscape.

## Limitations and Future Implications

Like any software, the specific tool at hand encounters certain limitations, which, although rare, could potentially impact its functionality:

- The tool retrieves data for websites ranking on the first page of search results. If there are numerous Google Ads or Google product listings, or if there is a substantial presence of rich results (structured data), the number of web pages appearing in the organic search results may be fewer than eight.
- The tool relies on web scraping techniques to access each competitor's website and identify the SEO techniques employed. If any of the competitor's websites are unavailable at the time of scraping or if the tool's access is blocked by the website's firewall, no results will be displayed for that particular website.
- The tool relies on the SERP API to gather ranking data. The data retrieval location is determined by the location of the API's web server. Consequently, search results and rankings from this specific location may slightly differ from what users observe from a different location.
- We must highlight that the accuracy of the model's predictions presented in **Machine Learning** section could be enhanced with a larger and more representative sample of websites.

### Usability Aspects:

- Regarding User Interface (UI) Design, the user interface has not been developed, which means that users will only be able to view the results either in Excel format or within the terminal.
- The software is user-friendly; however, it does require a certain level of background knowledge in executing Python code.
- As an open-source software, it can be easily customized and extended with additional features, such as a Usability and Security Testing function to predict Click-Through Rate [374].

### Security Aspects:

- Data Privacy is ensured since the software runs locally on the user's computer, and their data remains within their system without any exposure to the internet.
- Input Validation is implemented to ensure the security of the software. User-entered data is considered safe, and extreme validations are not applied to the input.

While all of the aforementioned limitations do not directly affect the functionality of the tool, they can potentially have a minor impact on the results.

Looking forward, there are several potential future implications and areas of development for this software. It is essential to integrate recently discovered SEO techniques into the software to ensure its relevance and effectiveness in a rapidly evolving SEO landscape. As we expand and enrich the ML dataset, we can anticipate achieving even more promising outcomes. The software's proficiency in identifying optimal SEO strategies and making precise predictions for backlinks and linking domains is directly influenced by the quality and quantity of accessible data. By consistently expanding and modifying the dataset:

- We could bolster the software's capability to enhance predictive accuracy: A larger dataset would result in heightened accuracy when forecasting the most effective quantities of backlinks and linking domains, thereby contributing to a more successful SEO strategy.
- We could employ unsupervised ML to uncover novel SEO techniques from well-established websites that have not yet been revealed to the broader public.
- We could adapt to evolving algorithms: Given the constant evolution of search engine algorithms, a robust dataset enables the software to adapt and remain current with these changes, ensuring the recommended SEO strategies maintain their efficacy.
- Semantic SEO: Leveraging semantic search principles to produce content aligned with user intent and context can enhance search visibility.

To sum up, the future potential of this software is substantial for amplifying its capabilities and delivering more influential outcomes. The primary catalyst for unlocking these advantages is the continuous expansion of the dataset, enabling the software to evolve, adapt, and provide increasingly valuable SEO guidance.

#### **4.7.5. Section Conclusion**

In conclusion, this study introduces an open-source SEO audit software developed in Python, heralding a significant advancement in the accessibility and affordability of sophisticated SEO analysis tools. The software stands shoulder to shoulder with its commercial counterparts, effectively obliterating financial barriers and extending its benefits to a broader user spectrum, including non-profit organizations. The orchestration of traditional programming techniques and cutting-edge ML algorithms seamlessly integrates into the software's implementation, resulting in a potent solution for discerning both on-page and off-page SEO strategies.

To elaborate further, this software not only addresses overarching enhancements for websites based on an array of on-page SEO performance metrics, but it also delves deeper, scrutinizing specific metrics and extending its analysis to competitors' webpages. In essence, this tool operates on a dual level. Firstly, it provides micro-level insights, outlining precise actions to be taken for each metric, and elucidating how these optimizations contribute to the overall enhancement of the website. Secondly, it offers tailored guidance that aligns with the unique needs and nature of a website, steering website owners towards areas of focus that match their specific requirements.

By harnessing a target website's URL and a specific keyword, the software not only identifies competitors within search results but also meticulously dissects their SEO methodologies, furnishing actionable recommendations to elevate the search ranking of the designated website. An especially compelling case study, centered around a fiercely competitive keyword, vividly underscores the software's efficacy in substantially augmenting organic traffic for the corresponding e-commerce platform. These revelations reverberate as a compelling call to action, underscoring the profound significance and value that germinate through the nurturing of open-source tools.

Finally, it is imperative to emphasize that preceding studies have managed to underscore the utility of certain SEO techniques and apply them in specific case studies. In contrast, our own research has aggregated all of these SEO techniques and has given rise to invaluable all-in-one SEO software. This software holds the potential to yield comprehensive results for the web developers who employ it.

## 4.8. SEO Techniques and Technologies Evaluation

### 4.8.1. Section Summary and Significance of the Study

Following our research on SEO, this study investigates the effectiveness of various SEO techniques and technologies, focusing on those that significantly improve a website's visibility and ranking on search engines. The primary aim is to identify strategies that increase organic traffic to websites within a constrained budget.

The study evaluates leading SEO practices, their impact on organic traffic, and their utilization by prominent websites and e-commerce platforms. This evaluation helps website owners make informed decisions on which SEO strategies to implement and prioritize, with the goal of enhancing their search engine rankings and attracting more visitors.

The research is structured as follows: **Section 4.8.2** introduces the key SEO techniques and technologies, including a PHP-based web crawler tool developed for this study. This tool examines the source code of active websites to identify the SEO strategies they employ. A selection of well-known websites and e-commerce platforms was analyzed using this tool. The section also outlines three research questions concerning the prevalence, importance, and prioritization of specific SEO techniques among top websites. Having the results from the own-developed tool, a figure is created containing all the SEO techniques used by the most well-known websites/E-commerce and the adoption rate is presented for each SEO technique that has been applied. A SEO technique with a higher adoption ratio is used by more top-listed websites, which makes it more important according to the research hypotheses of **Section 4.8.2**. Finally, the study concludes in **Section 4.8.5** summarizing the results, and presenting readers with the importance of this research.

### 4.8.2. Materials and Methods

A representative sample of top-listed websites and e-commerce platforms within their respective industry was collected from Alexa's website, totaling 2,000 entries. Among these, 1,689 entries represent online stores, while the remaining entries pertain to various websites. The list is provided in a CSV format [375].

The websites in this sample are ranked at the top in Alexa's ranking tool, determined by their daily traffic and unique visitors [376].

#### Research Questions

This sample forms the basis for our analysis, which aims to answer three critical questions about the importance, application, and prioritization of SEO techniques.:

1. Which SEO techniques are the most important?
2. Which specific SEO techniques have already been applied by the most well-known websites?
3. Which SEO technique should be implemented as a first priority?

#### Research Hypotheses

Six hypotheses have been formulated, forming the foundation for our research assumptions:

1. The larger the sample the more representative the results.
2. Top-listed websites have more correctly applied SEO techniques.
3. Top-listed websites have entire digital marketing team to support them, including specialized SEO experts who precisely understand the effectiveness of different SEO techniques and which ones must be implemented to achieve higher results in SERPs.
4. SEO techniques used by top-listed websites are also the most effective.
5. The most commonly used SEO techniques in our results should be prioritized.
6. The most commonly used SEO techniques in our results are the most effective.

### SEOmized PHP-Based Tool Implementation

Inspired by the SEO techniques and technologies thoroughly analyzed in **Sections 4.2** and **4.4**, we developed a PHP-based web crawler and web scraper tool, named SEOmized (Search Engine Optimized) [377], to automate the detection of SEO techniques in the websites' source code. SEOmized is responsible for extracting websites' HTML source code and identifying which SEO technique(s), among the predefined set of 17 SEO techniques presented in **Table 23**, are applied to the website. The source code of SEOmized is available in the GitHub repository [377].

### Limitations and Application Programming Interface (APIs)

To measure website speed and mobile-friendliness, SEOmized integrates two APIs into its functionality.

- Mobile-Friendly Test Tool API: Created by Google, this API scans the given URL against responsive techniques and returns a list of mobile usability issues [338].
- PageSpeed Insights API: Also created by Google, this API measures the performance of a web page, providing suggestions about the website's performance, including page speed, accessibility, and SEO [378].

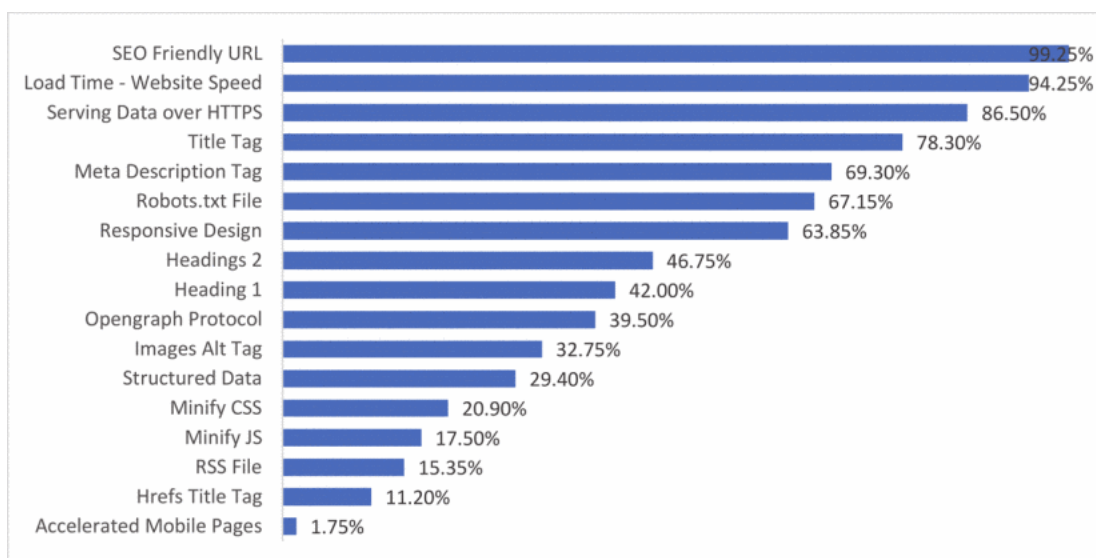
The API tools processed the URLs from our dataset, successfully retrieving data for most websites. However, the Mobile-Friendly Test Tool API failed to scan 24 websites, necessitating manual reviews of their responsive designs. On the other hand, the PageSpeed Insights API tool failed to deliver speed metrics for two websites, prompting us to use the Pingdom tool for those specific cases [342].

### Crawling Top-Listed Websites and Detecting SEO Techniques

The SEOmized tool was fueled by a list of 2,000 highly-ranked websites retrieved from Alexa's website. It systematically processes each website, identifying and cataloging the SEO techniques in use. This information is then stored in the SEOmized database, formatted in SQL [379]. The database contains records of the websites along with the specific SEO strategies they implement, marking the presence of applied SEO techniques.

#### 4.8.3. Research Findings

A list of the most commonly used SEO techniques among top-listed websites and their usage percentage is presented in **Figure 56** and **Table 22**.



**Figure 56.** SEO techniques usage score among top-listed websites (%).

**Table 22.** An illustration of the results in the SEOmized database

<b>seomized_id</b>	26	72
<b>website</b>	<a href="https://case-mate.com/">https://case-mate.com/</a>	<a href="https://www.batteryspace.com/">https://www.batteryspace.com/</a>
<b>title</b>	1	0
<b>meta_description</b>	1	1
<b>heading1</b>	0	0
<b>heading2</b>	1	0
<b>url</b>	1	1
<b>image_alt</b>	0	0
<b>href_title</b>	0	0
<b>rss</b>	0	0
<b>robots</b>	1	1
<b>responsive</b>	0	1
<b>https</b>	1	1
<b>structured_data</b>	1	0
<b>amp</b>	0	0
<b>minify_css</b>	0	0
<b>minify_js</b>	0	0
<b>og_tags</b>	0	1
<b>page_load_time</b>	1	1
<b>SUM</b>	<b>8</b>	<b>7</b>

The results enabled us to extract valuable information about how top-listed websites attempt to apply SEO techniques to increase their organic traffic.

To quantify the results of SEOmized, we introduced a corresponding metric, i.e., the SEOmized score. This score is equal to the ratio of the implemented SEO techniques and technologies detected in each website by SEOmized against the total ones supported by the tool.

**Table 23** provides a sample of results for three well-known e-commerce marketplace platforms (Ebay.com, Amazon.com, Walmart.com), as opposed to one less-known book store e-commerce (Bookhampton.com). The SEOmized score is calculated for each website, representing the percentage of total applied SEO techniques against the total ones supported by the tool..

**Table 23.** Website achievements against 17 SEO techniques

<b>SEO Techniques</b>	<b>Ebay.com</b>	<b>Amazon.com</b>	<b>Walmart.com</b>	<b>Bookhampton.com</b>
Title Tag	✓	✓	✓	✗
Heading 1	✓	✓	✗	✓
Headings 2	✓	✓	✓	✓
SEO Friendly URL	✓	✓	✓	✓
Meta Description Tag	✓	✓	✓	✓
Images Alt Tag	✗	✗	✗	✓
Hrefs Title Tag	✗	✗	✗	✓
RSS File	✗	✗	✗	✓
Robots.txt File	✓	✓	✓	✓
Responsive Design	✓	✓	✓	✓

Serving Data over HTTPS	✓	✓	✓	✓
Structured Data	✓	✗	✗	✓
Accelerated Mobile Pages	✓	✗	✗	✓
Load Time - Website Speed	1.3 seconds	2.1 seconds	1.2 seconds	3.9 seconds
Minify CSS	✗	✗	✗	✓
Minify JS	✗	✗	✗	✓
Opengraph Protocol	✓	✓	✓	✓
SEOmized Score	70,59%	58,82%	52,94%	94,12%

Following the hypotheses presented in the **Research Hypotheses** section, this research concludes as follows:

- 94.25% of websites have adopted technologies such as Apache Deflate and Gzip to reduce their load time to under 5 seconds. Specifically, the overall average load time among top-listed websites is 1.88 seconds.
- As discussed in **Section 4.4.3** the Chrome browser labels all websites without SSL certificates as 'not secure.' In response 86.50% of websites have installed SSL certificates to date.
- Despite search engines favorably evaluating websites using structured data, we noticed that 70.60% of the websites have not implemented structured data in their code.
- Moreover, despite the increasing use of mobile devices by visitors, only 63.85% of websites have passed the responsive-design test.
- **Table 23** showcases an online store of a bookstore among the results (Bookhampton.com). This online store utilizes 94.12% of the SEO techniques supported by SEOmized tool, unlike larger online stores that have applied fewer SEO techniques. This serves as a notable example of smaller websites, adhering to every SEO rule to ascend to the search results.

The presented results offer an initial exploration into the SEO techniques employed by prominent websites. By referring to **Figure 56** as a guide map, webmasters can initiate the application of SEO techniques with a higher adoption rate. Notably, certain techniques like Accelerated Mobile Pages and HREFs title tags seem less favored by top-listed websites, implying that their implementation might be more suitable for a later stage of SEO optimization. It's essential to note that these numbers may evolve daily; however, the overarching trend points towards a preference for a faster and more secure web environment.

#### 4.8.4. Discussion and Future Work

The data collected from this research contributes to a clearer understanding of how well-known websites construct their SEO strategy. Although the majority of the websites have applied SEO techniques for security and speed, they have failed to use certain techniques that the search engines consider important, such as the AMP, RSS, Minified files, and Structured data. An assumption could be that these techniques are challenging to apply to overcrowded websites. For instance, creating an AMP alternative page requires website owner to maintain two versions of their website. Therefore, it's convenient for top-listed websites not to use those SEO techniques that demand a lot of effort to maintain. Instead, they purposely sacrifice the extra traffic that these techniques could potentially deliver, and attempt to gain it by using different methods, such as allocating resources for customer support and satisfaction.

With more techniques and technologies that promote SEO being created daily, SEO has become an infinite marketing industry that generates income for developers and yields results for website

owners. The SEOmized tool created for the needs of this study can be extended with additional SEO techniques in the future, providing more alternatives to webmasters and leading to even greater results.

#### **4.8.5. Section Conclusion**

Since the mid-1990s, webmasters have recognized the crucial role of high search engine rankings for their websites [283]. In pursuit of this goal, they have continuously sought to enhance their sites for search engines, either by adopting existing technologies or innovating new ones. Over the years, SEO strategies have been refined through trial, observation, and the accumulation of experience, making SEO a cornerstone of website promotion efforts. This study aims to provide Webmasters with a strategic roadmap for employing effective SEO techniques on their websites. Three main questions were addressed throughout this research: What are the most critical SEO techniques? Which SEO techniques are most commonly used by leading websites? And which SEO technique should be prioritized for implementation?

To answer the above questions, we developed a SEO tool designed to analyze a website and pinpoint the specific SEO techniques it employs. This tool was given a list of 2,000 top-listed websites as input and tasked with scanning them to reveal the SEO techniques each one follows. The findings were compiled into a database, culminating in the creation of a figure **Figure 56** that ranks the prevalence of SEO techniques among these top sites. This figure can serve as a guide-map for Webmasters, answering the research questions and guiding them in adopting the SEO tactics proven successful by the most renowned websites. By adhering to the SEO blueprint used by these leading sites, any Webmaster, regardless of their SEO expertise, can implement effective SEO strategies. This approach will help improve their site's ranking in SERPs and increase their web traffic.

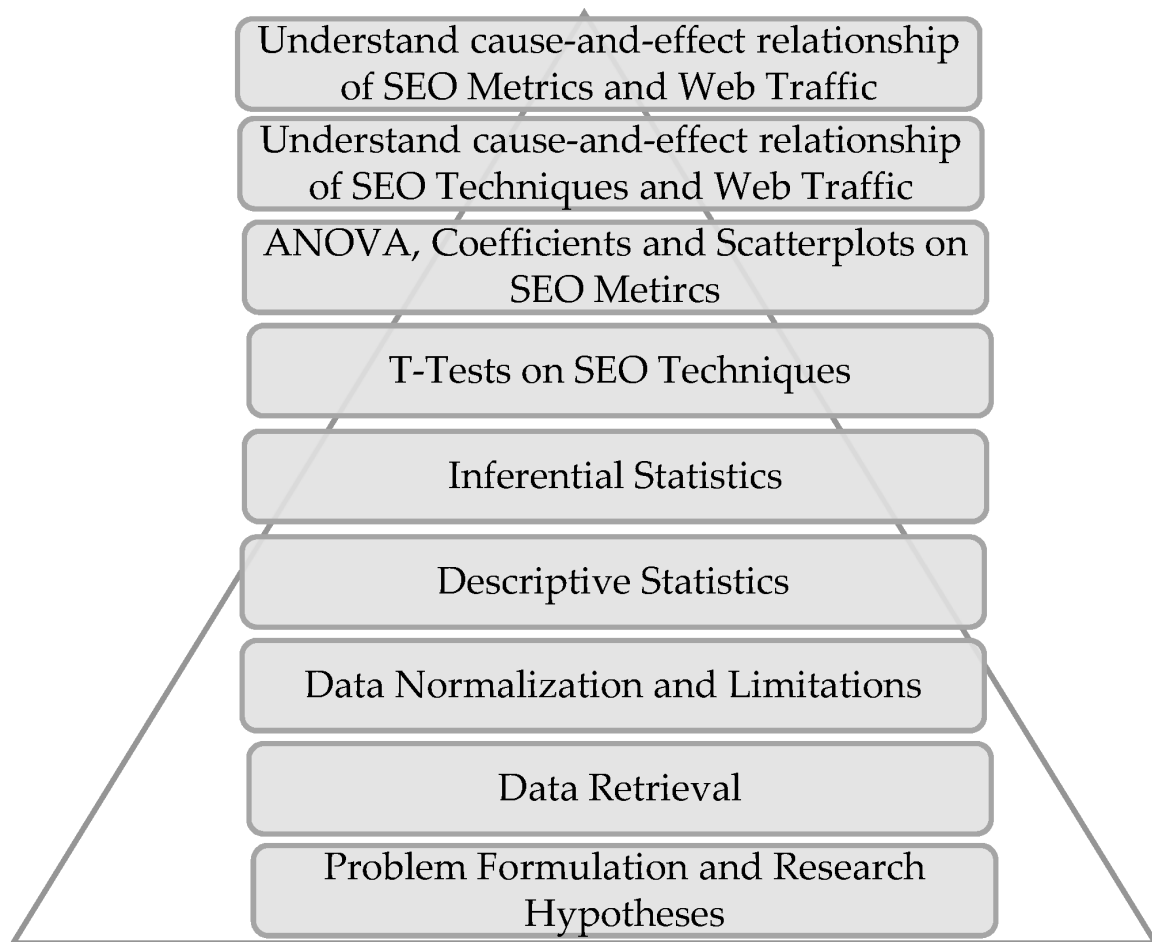
### ***4.9. Airlines' Sustainability Study Based on Search Engine Optimization Techniques and Technologies***

#### **4.9.1. Section Summary and Significance of the Study**

Digital marketing, especially SEO, is an integral part of websites today. Airlines in the COVID-19 era have to use every possible means to survive despite the adverse conditions for both entrepreneurship and travel. Many of them have allocated resources and money to develop SEO strategies by applying SEO techniques to their websites to gain more visitors and bookings through their service-oriented e-commerce platforms. In continuation of our research on SEO, our focus is on analyzing the online presence of airlines, specifically examining the implemented SEO techniques and their impact on airlines' website traffic. In the first phase of the research, we gathered web data from 243 airline firms during a one-year observation period (December 2020–December 2021) using our own-developed PHP-based tool , which was referenced in **Section 4.8** and upgraded to meet the requirements of this specific case study. Furthermore, we proceeded to create an exploratory model using fuzzy cognitive mapping. From the technical SEO point of view and the descriptive analysis, we conclude that the traffic on airlines' websites and, consequently, their sustainability are inseparably linked to the corresponding SEO techniques and technologies used.

#### **4.9.2. Materials and Methods**

The purpose of this study is the proposal of an innovative methodology for deploying an efficient framework for understanding the strong connection between SEO and web traffic, providing valuable insights for airline firms' websites. To reach the desired results, we follow nine research stages, presented in **Figure 57**.



**Figure 57.** Overall step-by-step representation of the proposed methodology, reflecting the stages for comparing SEO techniques and SEO metrics to Web traffic.

### **Websites' Traffic Sources and Key Performance Indicators**

Airlines' online presence, as well as their bookings, are directly affected by various factors, which should be analyzed and taken into consideration over our estimation. To obtain a more specific vision of the airlines' online presence, we use the organic web traffic source, which indicates the main source of websites' traffic. Organic traffic, as mentioned in previous sections, is the traffic coming from users who use search engines (e.g., Google, Yahoo, etc.). It is an unpaid form of web traffic, whose role in business success is vast. The majority of websites rely on organic traffic results, since it represents over 60% of a website's total traffic [380]. Organic traffic, as an unpaid source of traffic, has many benefits for businesses, such as relatively small investment in contrast with PPC campaigns, long-lasting results enhanced year by year with the appropriate effort, and major increase in brand awareness [381].

Four more factors that search engines take into consideration while evaluating a website, before ranking it in the SERPs, are the organic keywords, the backlinks, the website speed, and the DA [382]. Organic keywords are the keywords for which websites are ranked in the organic SERPs of search engines [383]. Backlinks are links from third pages to the target page that act as an up-vote on that page [274]. Website speed is the time in seconds it takes for a web page to be fully downloaded and loaded from the webserver to the user's browser. DA is a search engine ranking score developed by Moz that predicts the strength of a website in terms of its ability to rank in SERPs. DA score ranges from 1 to 100, with higher scores corresponding to greater likelihood of ranking [367]. DA as a metric is not visible to search engines. It tries to simulate how search engines rank web pages. In most cases, the DA score is

representative. DA is important to compare websites with each other. Websites with a higher DA are more likely to appear higher in SERPs than others with a lower DA, respectively.

The web traffic metrics proposed as KPIs in this study are presented in **Table 24**.

A performance indicator or key performance indicator (KPI) is a type of performance measurement that evaluates a business' success in certain activities that it takes over [384,385]. KPIs examined for the purposes of this study alongside the web metrics are included in **Table 24**. It is valuable to examine website traffic on a monthly basis to export more accurate results concerning traffic and ranking performance.

The main objective of our KPIs is to achieve higher results in organic traffic as well as higher positions in search results. In our research, we have five key performance indicators, i.e., Organic Traffic, DA, Website Speed, Organic Keywords, and Backlinks. Each one should be accomplished to gain higher organic traffic and potential customers.

**Table 24.** Description of the examined web analytics metrics and suggested KPIs and performance measurements.

KPIs (Unit)	Description
<b>Organic Traffic/month (number of visitors per month)</b>	Organic search refers to the non-paid search results from a search engine. These results can't be bought or influenced by advertisers, so they are the ones the search engine considers most relevant to the user's search query [386].
<b>DA (score in the range (0–100) integer)</b>	DA score can be used when comparing websites or tracking the "ranking strength" of a website over time [384]. Websites with higher DA are more likely to rank higher in SERPs than those with lower DA.
<b>Website Speed (seconds)</b>	Website speed refers to how quickly a browser can load fully functional web pages from a given website.
<b>Organic Keywords (number of keywords)</b>	Organic keywords are keywords used in SEO to attract "free" traffic. When a user types a keyword on the search engines gets as a result websites which rank for this specific keyword [383].
<b>Backlinks (number of backlinks)</b>	Backlinks are the number of external pages that show the target page with a link. Backlinks help secure a higher SERP ranking [277]. Backlinks affect mostly the position of a website followed by relevant content [387].

### Sample Selection and Data Retrieval

For the purpose of this study, we retrieved data from the International Air Transport Association (IATA), which supports aviation with global standards for airline safety, security, efficiency, and sustainability [388]. The sample is based on 243 IATA members. Some of the most well-known airlines in our sample are Air Canada, Air France, American Airlines, British Airways, China Airlines, and Emirates. For the above airline firms, data were collected concerning their websites' organic, direct, and referral traffic, as well as their websites' session duration and the bounce rate. The sample was considered representative, providing knowledge about the SEO applied by airlines as well as how this translated into traffic. Data were collected on a monthly basis from the airlines' websites, for more accurate examination and comprehension of web metrics' variance. The testing period

extended to 12 months, which is limited to specific dates from December 2020 to December 2021. Since this period coincides with the COVID-19 transportation restrictions, the results will provide a clear view of the decline in bookings during this period.

To collect the sample, we created a PHP-based web crawler tool, which undertook to extract information, such as legal name, ICAO code, region, and website from each airline listed as a member in IATA's website [388,389]. After the data extraction, we stored the entire dataset in a MySQL database, which we used to conduct the SEO checks as a next step [389].

The software tool created for the purpose of this research draws the dataset from IATA's website and stores each result in a MySQL database. It then runs through the dataset and extracts the source code from each airline's website via cURL. CURL is a computer software project providing a library (libcurl) and a command-line tool (curl) for transferring data using various network protocols [390]. For each of the 15 SEO techniques presented in **Section 4.2**, a separate function was created, which searches the source code of the targeted website to find out if the SEO technique has been applied. Finally, the results from the checks are stored back in the database. A comprehensive example of the generated code and its main functions is shown in **Figure 58**.

```
require ('seo-techniques-functions.php'); // ToDo! Get SEO techniques' functions
$retrieve = retrieve_data_from_iata(); // ToDo! Retrieve data from IATA's website
$websites = create_dataset_array($retrieve); /* ToDo! Create an array with our dataset */
foreach($websites AS $id => $website) { // ToDo! Run through our dataset
    $curl_results = do_curl_website($website); // ToDo! Do curl on current website
    $seo_checks = do_seo_checks($curl_results); /* ToDo! Do SEO checks using the
    getElementsByTagName */
    $add_results_to_db = add_results_to_db($seo_checks,$website); /* ToDo! Add SEO
    results to DB */
}
```

**Figure 58.** PHP-based SEO software.

To expand our research and extend our results, four third-party API tools were used, i.e., Mobile-Friendly Test Tool API, PageSpeed Insights API, Mozscape API, and Ubersuggest Traffic Analytic tool. Each of these tools obtains the website's URL and returns results based on its own measurements.

- Mobile-Friendly Test Tool API, created by Google, checks the given URL against responsive techniques and returns a list of any mobile-usability issues that can affect a user that visits the page on a mobile device [334].
- PageSpeed Insights API, created by Google, measures the performance of a web page returning suggestions on how to improve the page's performance, such as page speed, accessibility, and SEO [378].
- Mozscape API, created by MOZ, takes as an input the website's URL returning accurate metrics such as DA [367].
- Ubersuggest, created by Neil Patel, obtains and returns a 360° overview of any website's metric and its sources [391].

After conducting SEO evaluations on 243 airline websites, the exported results imported to IBM SPSS 27 (IBM, Armonk, NY, USA). Organizations use IBM SPSS Statistics to understand data, analyze trends, forecast and plan to validate assumptions, and drive accurate conclusions [392]. Descriptive analysis in different correlations was conducted, confirming the research hypotheses that follow. Finally, a fuzzy cognitive map (FCM) was created to graphically represent the relationship between SEO techniques, traffic, and backlinks, showing the direct correlation of SEO techniques with traffic and, consequently, the bookings and sustainability of airlines.

## Problem Formulation and Research Hypotheses

The rising competition, the COVID-19 era, and travel restrictions around the world tend to increase companies' efforts for efficiency in digital marketing campaigns. Airlines use every available tool at their disposal, from copywriting, email marketing, social media marketing to SEO to reach their potential customers through numerous methods. By understanding the consumer habits of their audience, airlines develop marketing strategies to target this audience more effectively.

Airline websites are the main source of bookings and companies dedicate resources to organize and maximize their website's efficiency and effectiveness. Their position in SERPs plays a crucial role in gaining organic visitors. The web page title, the meta description, and even the rich results must be applied with attention to detail to win the coveted click. Content, user interface, and website speed will pique the visitor's interest converting him/her to a loyal client. Knowing that traffic and conversion to client is the main point of airlines, we define five hypotheses on which we base our research, attempting to confirm them, aiming to extend the practical knowledge over the significance and impact of SEO to airline website traffic.

- **Hypothesis 1 (H1).** SEO techniques impact airlines' website traffic.

The main goal of this study is to acknowledge whether SEO impacts airlines' website traffic. By knowing which SEO techniques airlines implement on their websites, comparing the findings with the results in traffic provides us with valuable information and potential benefits from using SEO techniques. As mentioned in **Sample Selection and Data Retrieval** section, a dataset was collected from the IATA website [388], consisting of 243 airline firms' websites. Scanning and performing SEO tests on each website source code, using our own developed software [389], we gather information regarding the SEO techniques each website has implemented. Using third-party software tools described in **Sample Selection and Data Retrieval** section, we scan each website gathering information concerning their DA, website speed, organic keywords, organic traffic, and backlinks. All data collected are stored in an excel sheet. Using the IBM SPSS 27 software descriptive analysis was conducted in order to find connections/patterns between SEO techniques used by websites and their organic traffic. The ultimate goal of this research is to uncover which SEO techniques used by the most well-known airlines are more effective in organic traffic, assuming that the websites with the highest organic traffic applied the appropriate SEO techniques in their source code.

- **Hypothesis 2 (H2).** On-page SEO techniques impact airlines' website ranking.

Hypothesis 2 is a segment of Hypothesis 1, targeting the on-page SEO techniques. Our second hypothesis is based on on-page SEO techniques and how they affect website ranking. Targeting multiple keywords, as well as the implementation of on-page SEO techniques, is assumed to offer an increase in organic results.

- **Hypothesis 3 (H3).** Off-page SEO techniques—backlinks impact airlines' website ranking.

Hypothesis 3 is a segment of Hypothesis 1, targeting off-page SEO techniques. Hypothesis 3 focuses on the implementation of off-page SEO techniques and how they affect the traffic of airlines' websites. We assume that the creation of backlinks to third-party websites will lead to strengthening website's DA and consequently will lead to higher positions in the SERPs.

- **Hypothesis 4 (H4).** Airlines' organic keywords impact airlines' website ranking.

In Hypothesis 4 we assume that organic keywords are valuable for airlines' website leading to higher rankings on SERPs and increased Web traffic.

- **Hypothesis 5 (H5).** Airlines' backlinks impact the airlines' DA.

In Hypothesis 5 we assume that the backlinks are integrally linked with DA.

### 4.9.3. Results

#### Data Normalization

In this section, we examine the collected data. As mentioned in **Sample Selection and Data Retrieval** section, our tool performed SEO checks on 243 websites. The results were stored in excel so that we can use them in further data analysis [393].

Collected data are separated into three groups.

- i. **Group 1:** Metrics are checks performed by third-party tools and are not SEO techniques. The contents of the Group Metrics are as follows: DA (integer 0 to 100), speed\_test (decimal), number of keywords (integer), and number of backlinks (integer).
- ii. **Group 2:** SEO techniques are checks performed by our tool regarding the existence (or not) of SEO techniques on each website in the dataset. All metrics returned 1 if the SEO technique was applied to the website and 0 otherwise. The contents of the Group SEO techniques are as follows: images\_alt, links\_title, rss, sitemap, robots, heading1, heading2, web\_ssl, meta\_description, opengraph, url\_seo\_friendly, amp, minified\_css, minified\_js, title, structured\_data, and responsive\_test.
- iii. **Group 3:** Web Traffic (organic traffic) is the monthly statistical data collected for each website. All web traffic data are integers and refer to website data for the last twelve months. The contents of the Group Web Traffic are as follows: traffic\_1, traffic\_2, traffic\_3, traffic\_4, traffic\_5, traffic\_6, traffic\_7, traffic\_8, traffic\_9, traffic\_10, traffic\_11, and traffic\_12.

#### Limitations

During the examination, the responsive\_test technique found that the Mobile-Friendly Test Tool API created by Google considers non-responsive pages that have even one non-responsive element. Manual checks revealed that these specific websites partially adopt responsive techniques and have several elements in their source code that are non-responsive. These pages are marked with the number 0, which means that they do not follow that specific SEO technique. Further, in 17 out of the 243 websites, the speed measurements were performed manually with the Pingdom website speed test by Solar Winds, because the PageSpeed Insights API by Google could not operate a speed test on these pages.

The AMP and Sitemap SEO technique tests found that none of the websites applied these techniques, although they are suggested by Google on Webmaster Guidelines [271]. Consequently, t-test cannot be performed on these techniques, excluding them from our final results.

We collected the monthly organic traffic for 12 months to have a more representative view of the website traffic. For the data analysis that follows, we calculated the mean value of 12 months of organic traffic for each website and saved it in a new column, called Web Traffic. All the following comparisons regarding web traffic refer to the mean value of organic traffic for each website.

Raw data collected from multiple websites do not convey any significant trends or the behavior of the individuals before the corresponding analysis. Data analysis refers to the process of converting raw data into meaningful information, using mathematical, statistical, or computational algorithms for better comprehension [394]. For the quantitative data analysis, some major steps need to be considered to firmly execute the assessment and generate the intended results. The procedure consists of two major steps: (1) reviewing the data by descriptive analysis and (2) conducting inferential and descriptive statistics to answer the research questions [395]. The study is presented in the following sections of the chapter and each section is intended to answer the research questions, as the analysis must be aligned with the research objectives.

#### Descriptive Statistics

Descriptive analysis helps to describe, demonstrate, or summarize the collected data in a constructive manner so that the trends and patterns can be easily observed and analyzed [396]. Mean,

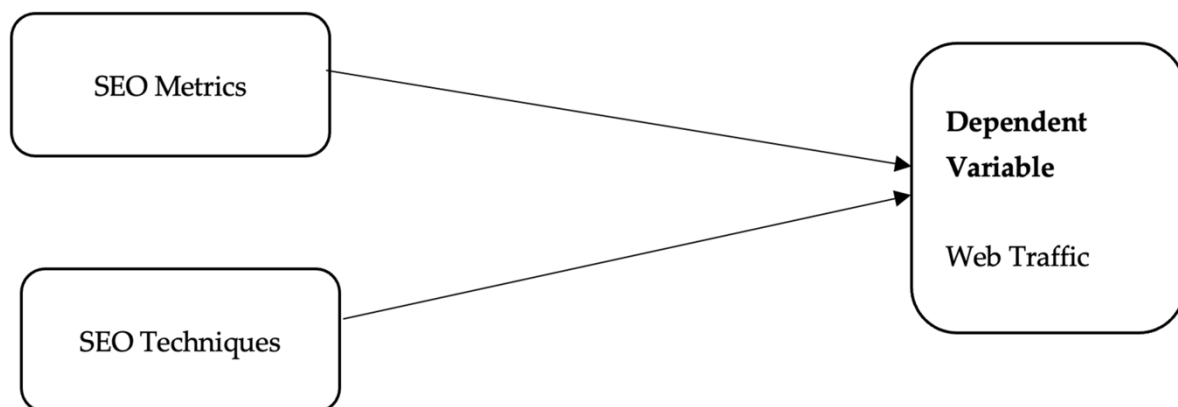
median, mode, standard deviation, skewness, and kurtosis are some of the important measures. **Table 25** indicates that the maximum DA, speed test, keywords, and backlinks in this study are 39,833,333.33, 88, 17.67, 2,272,818, and 117,837,438, respectively. According to Tabachnick, Fidell, and Ullman 2017, the values of skewness or kurtosis within -1.5 and +1.5 is considered right-skewed distribution of the data [397].

**Table 25.** Descriptive statistics on SEO metrics.

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Web Traffic	243	0.00	39833333.33	1052702.396	4047257.865	7.984	0.156	70.937	0.311
DA	243	10	88	56.23	14.919	-0.102	0.156	-0.641	0.311
Website Speed	243	0.20	17.67	3.8135	2.56189	2.148	0.157	7.053	0.312
Organic Keywords	243	0	2272818	147851.87	318947.390	3.846	0.156	17.216	0.311
Backlinks	243	1076	117837438	2587599.06	9450892.729	9.201	0.156	99.836	0.311

### Inferential Statistics

Inferential statistics refers to the procedure of using data analysis to infer attributes of an underlying distribution of probability. Inferential statistics deals with the information acquired from the sample of the population to draw conclusive statements about the entire population [398]. Inferential statistics make use of statistical models to compare the sample data to that of earlier studies. It varies from descriptive statistics in that it allows one to draw conclusions based on extrapolations rather than merely reporting the data that have been seen, like descriptive statistics does [399]. The theoretical framework of this study is given in **Figure 59**.



**Figure 59.** Testing research model.

**Impact of Group: SEO Techniques on Web Traffic**

At this point, 15 different t-tests were performed for each of the 15 SEO techniques: Group 2 using SPSS 27 by IBM software. After, the t-test mean values were observed. The mean value is equal to the sum of all the values in the dataset divided by the total number of values. The “YES” mean reflects the average web traffic from the websites that applied the technique. The “NO” mean reflects the average web traffic from the websites that did not apply the technique. If the “YES” mean is greater than the “NO” mean, the websites that implemented this technique in their source code have greater traffic and the technique is valuable for the website. Another component to consider when comparing Yes or No is the p-value. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis [400]. A p-value less than 0.05 (typically  $\leq 0.05$ ) is statistically significant. A p-value higher than 0.05 ( $>0.05$ ) is not statistically significant and indicates strong evidence for the null hypothesis [401]. The results from 2 of 15 t-tests are presented in detail below in **Table 26** and **Table 27** and the rest are presented briefly in **Table 28** for the sake of brevity.

**Table 26.** Independent t-test on the meta description technique.

Technique	Mean (number of visitors per month)		Standard Deviation (number of visitors per month)		t-value	p-value
	Yes (n=169)	No (n=74)	Yes (n=169)	No (n=74)		
Meta Description	1268843.04	559083.87	4712815.92	1682786.61	-0.260	0.00

**Impact of meta description technique on the web traffic**

**H0:** There is no difference in website traffic due to the meta description technique (null hypothesis).

**H1:** There is a difference in website traffic due to the meta description technique.

In this study, independent samples t-test is applied to make a comparison of the mean web traffic between websites that implement the meta description technique (n = 169) and websites that do not implement the meta description technique (n = 74). The t-test was statistically significant, as the mean web traffic of meta description technique implemented websites (M = 1,268,843, SD = 4,712,815) was higher than the meta description technique non-implemented websites (M = 559,083, SD = 1,682,786),  $t = -0.260$ ,  $p < 0.05$ , two-tailed. Therefore, the null hypothesis H0 that there is no difference in website traffic due to the meta description technique can be rejected. Consequently, it can be concluded that websites that implement a meta description technique have greater traffic and, by extension, this technique is valuable for the websites (**Table 26**).

**Impact of image alt technique on web traffic**

**H0:** There is no difference in website traffic due to the image alt technique (null hypothesis).

**H1:** There is a difference in website traffic due to the image alt technique.

In this study, independent samples t-test is applied to make a comparison of the mean web traffic between websites that implement image alt technique (n = 56) and websites that did not implement the image alt technique (n = 187). The t-test was not statistically significant since the p-value is higher than 0.05 and the mean web traffic of image-alt-technique-implemented websites (M = 1,156,760, SD = 4,532,209) was not significantly lower than the image alt technique non-implemented

websites ( $M = 705,223$ ,  $SD = 1,564,573$ ),  $t = 0.732$ ,  $p > 0.05$ , two-tailed. Therefore, the null hypothesis  $H_0$  that there is no difference in website traffic due to the image alt technique cannot be rejected (**Table 27**).

**Table 27.** Independent t-test on the image alt technique.

Technique	Mean (number of visitors per month)		Standard Deviation (number of visitors per month)		t-value	p-value
	Yes (n=56)	No (n=187)	Yes (n=56)	No (n=187)		
Image Alt	1156760.24	705223.51	4532209.344	1564573.19	0.732	0.46

**Table 28** presents the 15 t-tests performed for the 15 corresponding SEO techniques. In the last column of Table 5, SEO techniques that resulted in a positive impact on the traffic are marked as 1, while the techniques that did not have a significant contribution to the traffic are marked as 0.

**Table 28.** T-test on 15 SEO techniques

Technique	Mean (number of visitors per month)		Standard Deviation (number of visitors per month)		t-value	p-value	Valuable
	Yes	No	Yes	No			
Meta Description	1268843.04	559083.87	4712815.92	1682786.61	-0.260	0.00	1
Image Alt	1156760.24	705223.51	4532209.344	1564573.19	0.732	0.46	0
SEO friendly URL	1055012.32	493699.58	4055485.45	0	-0.138	0.03	1
Minified CSS	1767599.41	911835.98	6400155.29	3405036.20	1.244	0.00	1
Title	1110913.20	738670.37	4339095.14	1791340.36	-0.520	0.00	1
Structured data	1543444.60	931904.31	5804001.52	3492528.17	-0.938	0.02	1
Responsive	1166677.80	881151.78	4809189.46	2511798.05	-0.538	0.00	1
Heading1	1537246.33	602768.73	5601187.82	1474452.38	1.807	0.01	1
Heading2	1171771.22	863965.62	4701758.10	2717534.64	0.577	0.01	1
SSL	1078881.88	18612.63	4094980.17	18023.60	-0.633	0.00	1
Open graph	859248.90	1135773.60	2728729.00	4501161.79	0.487	0.04	0
Link Title	1072927.18	1048835.89	2082867.15	4326089.90	0.775	0.00	1
RSS	72420.45	1165120.05	172825.78	4259192.27	1.28	0.03	0
Robots	526336.78	1433477.51	1308707.77	5170205.59	1.73	0.01	0
Minified JS	772538.98	1075115.46	1727237.15	4178916.51	0.305	0.00	0

#### *Impact of Group: SEO Metrics on the Web Traffic*

Regression is a statistical technique to formulate the model and analyze the relationship between the dependent and independent variables. It aims to study the degree of relationship between two or more variables. This is achieved with the help of hypothesis testing.

In this part of our research, three different tests (one-way analysis of variance (ANOVA), coefficients, and scatterplots) were performed for each of the four metrics: Group 1 using the SPSS 27 by IBM software. The analysis is intended to determine the relationship between metrics and web traffic.

The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups or not. Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. A simple scatterplot can be used to (a) determine whether a relationship is linear, (b) detect outliers, and (c) graphically present a relationship between two continuous variables [402].

The null and alternate hypotheses remain the same as the techniques.

- H0: (Metric) does not have a positive impact on the web traffic.
- H1: (Metric) has a positive impact on the web traffic.

The model summary table reports the strength of the relationship between the model and the dependent variable (**Table 29**).

- DA: The regression model summary is explained by 15.9% variability in the web traffic,  $R^2 = 0.159$ , adjusted  $R^2 = 0.155$ .
- Keywords: The regression model summary is explained by 65.2% variability in the web traffic,  $R^2 = 0.652$ , adjusted  $R^2 = 0.650$ .
- Website Speed: The regression model summary is explained by only 0.1% variability in the web traffic,  $R^2 = 0.001$ , adjusted  $R^2 = -0.003$ .
- Backlinks: The regression model summary is explained by 86.9% variability in the web traffic,  $R^2 = 0.869$ , adjusted  $R^2 = 0.869$ .

**Table 29.** Model summary. Strength of the relationship between the model and the dependent variable.

Metric	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
DA	1	0.399	0.159	0.155	9223.372036854777000
Website Speed	1	0.034	0.001	-0.003	9223.372036854777000
Organic Keywords	1	0.807	0.652	0.650	9223.372036854777000
Backlinks	1	0.932	0.869	0.869	9223.372036854777000

The ANOVA table represents the overall significance of the model, which is determined by the web traffic (**Table 30**). The F ratio is the ratio of two mean square values. Degrees of freedom (DF) of error was calculated by total sample ( $n = 243$ )  $- 2 = 241$  and total DF was calculated by total sample ( $n = 243$ )  $- 1 = 242$ .

- DA: The result shows  $F(1, 241) = 45.524$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this model is related to the dependent variable.
- Keywords: The result shows  $F(1, 241) = 450.989$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this is related to the dependent variable.

- Website Speed: The result shows  $F(1, 239) = 0.277$ ,  $p > 0.05$ , which suggests the model is not statistically significant. Therefore, the independent variable in this is not related to the dependent variable.
- Backlinks: The result shows  $F(1, 241) = 1604.989$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this model is related to the dependent variable.

**Table 30.** ANOVA results. Overall significance of the model.

Model		Sum of Squares	df	Mean Square	F	p-value
DA	Regression	629824572758098.500	1	629824572758098.500	45.524	0.000
	Residual	3334207114340703.000	241	13834884291870.137		
	Total	3964031687098801.500	242			
Website Speed	Regression	4585155651162.500	1	4585155651162.500	0.277	0.599
	Residual	3957236139014683.500	241	16557473384998.676		
	Total	3961821294665846.000	242			
Organic Keywords	Regression	2583473225519711.000	1	2583473225519711.000	450.989	0.000
	Residual	1380558461579090.200	241	5728458346801.204		
	Total	3964031687098801.000	242			
Backlinks	Regression	3446514090608055.000	1	3446514090608055.000	1604.989	0.000
	Residual	517517596490746.700	241	2147375919048.742		
	Total	3964031687098801.500	242			

The coefficient table represents how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant [403] (Table 31). When interpreting the coefficient table, the B variable is important, representing the increase in the dependent variable as soon as the independent increases by one unit. On the other hand, the Beta variable compares the strength of the effect of each individual independent variable to the dependent variable. The higher the value of the beta coefficient, the stronger the effect, based on absolute numbers [404].

- DA: When the beta value of DA is increased by one unit, web traffic will have increased by 108,135 visitors per month. Therefore, DA does not have a positive impact on the web traffic; thus, the null hypothesis is rejected. It can be concluded that the DA scores have a significant positive effect on the web traffic.
- Keywords: When the beta value of keywords is increased by one unit, the web traffic will have increased by 10.24 visitors per month. Therefore, keywords do not have a positive impact on the web traffic; thus, the null hypothesis is rejected. It can be concluded that the keywords have a significant positive effect on the web traffic.
- Website speed: Since the p value of the following table is greater than 0.05, the impact of speed is not statistically significant. Therefore, the null hypothesis that speed test does not have positive impact on the web traffic is rejected.
- Backlinks: When the beta value of backlinks is increased by one unit, the web traffic will have increased by 0.399 visitors per month. Therefore, backlinks do not have a positive impact on the web traffic; thus, the null hypothesis is rejected. It can be concluded that the backlinks have a significant positive effect on the web traffic.

**Table 31.** Coefficient results. How much dependent variable is expected to increase when that independent variable increases by one.

Model	B	Std. Error	Beta	t	p-value
(Constant)	5027333.372	932178.327		-5.393	0.000
DA	108135.014	16026.707	0.399	6.747	0.000

(Constant)	1267137.643	470709.928		2.692	0.008
Website Speed	-53952.480	102525.384	-0.034	-0.526	0.599
(Constant)	-461911.986	169294.489		-2.728	0.007
Organic Keywords	10.244	0.482	0.807	21.237	0.000
(Constant)	19449.818	97478.822		0.200	0.842
Backlinks	0.399	0.010	0.932	40.062	0.000

Scatterplots are essential for identifying trends and patterns. In a scatterplot, each observation (or point) has two coordinates [405]. The strength of the link between the variables is determined by calculating the correlation coefficient. The plot shows the first variable's value on the x-axis and the second variable's value on the y-axis for each data point [406]. The relationship between two quantitative variables is shown in a scatter plot (**Figure 60, Figure 61, Figure 62, Figure 63**).

- DA: The following graph indicates the coordinates of DA and web traffic. The web traffic has been gradually increased because of the increase in DA scores, indicating a linear relationship between DA and the web traffic.
- Keywords: The following graph indicates that the more keywords used on a website, the more traffic it will get. Therefore, there is a strong positive linear relationship between keywords and web traffic.
- Website speed: The following shows the relationship between speed test and web traffic. It shows that an increase in speed test does not result in a proportional increase or decrease in web traffic. Therefore, there is a constant and nonlinear relationship between these two variables.
- Backlinks: Like keywords, the following graph indicates that the more backlinks used on a website, the more traffic it will get. Therefore, it can be concluded that there is a strong positive linear relationship between keywords and web traffic.

We used exactly the same data analysis methodology as the SEO metrics to identify the correlation between backlinks and DA. The findings of the regression model found that backlinks have a strong connection with DA. When the beta value of backlinks is increased by one unit, on average, the DA will have increased by 6.004 units.

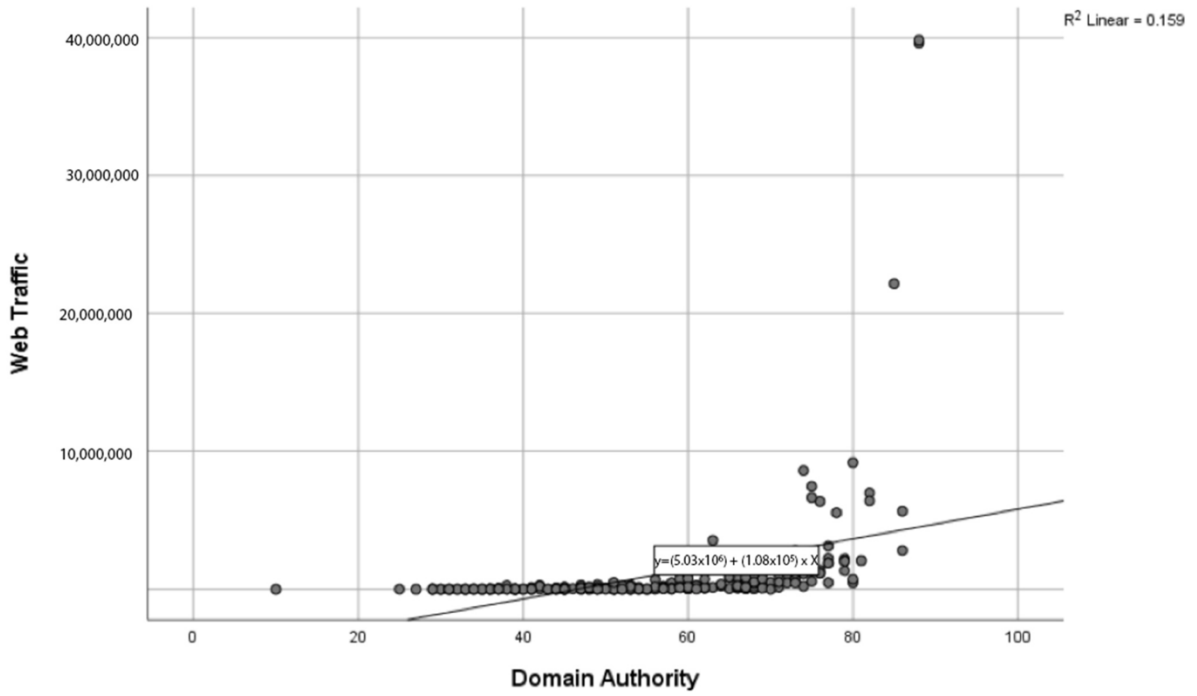


Figure 60. Scatterplot of DA by web traffic.

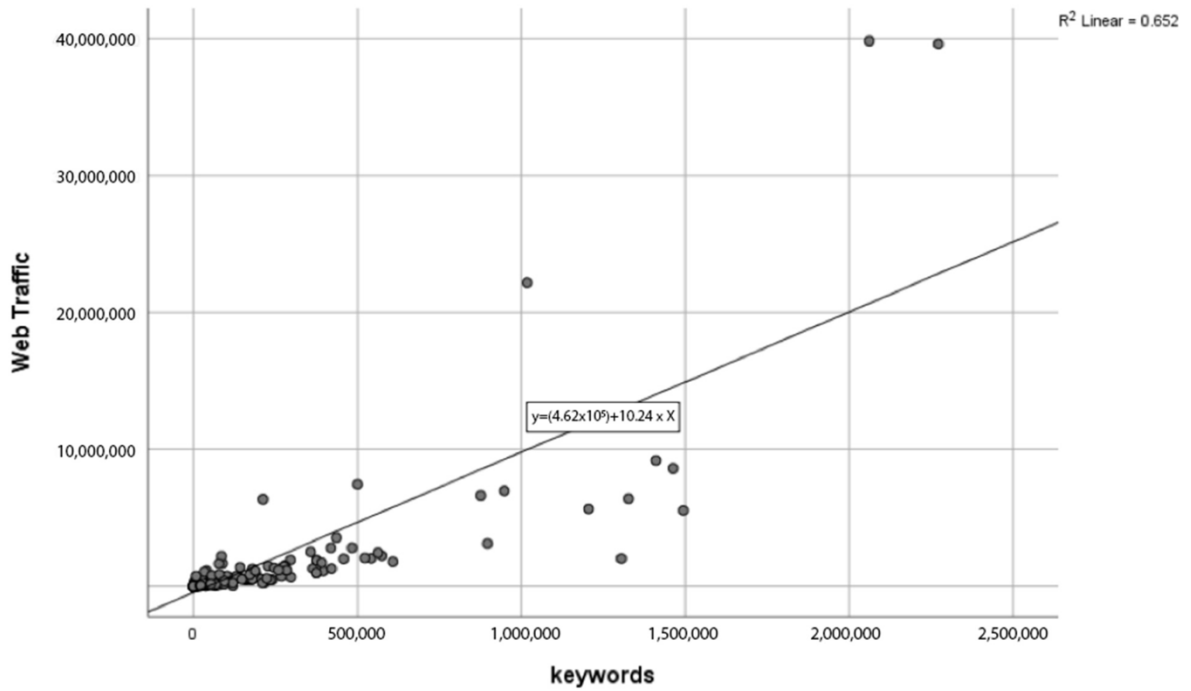


Figure 61. Scatterplot of Keywords by web traffic.

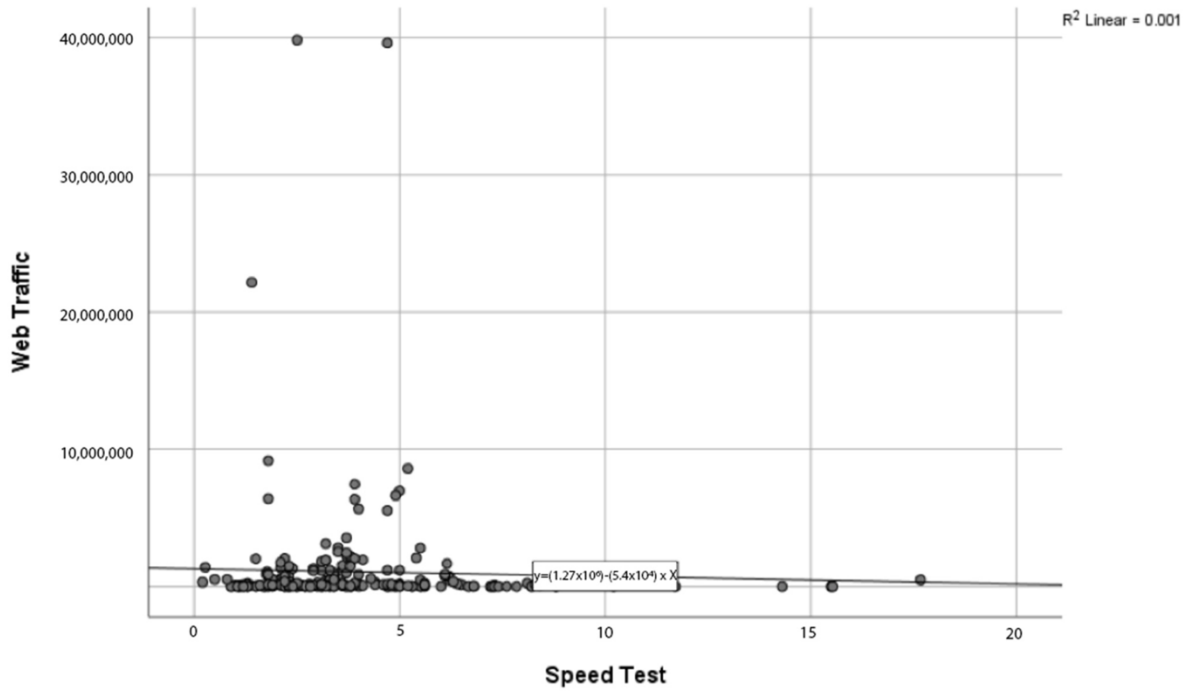


Figure 62. Scatterplot of Website Speed by web traffic.

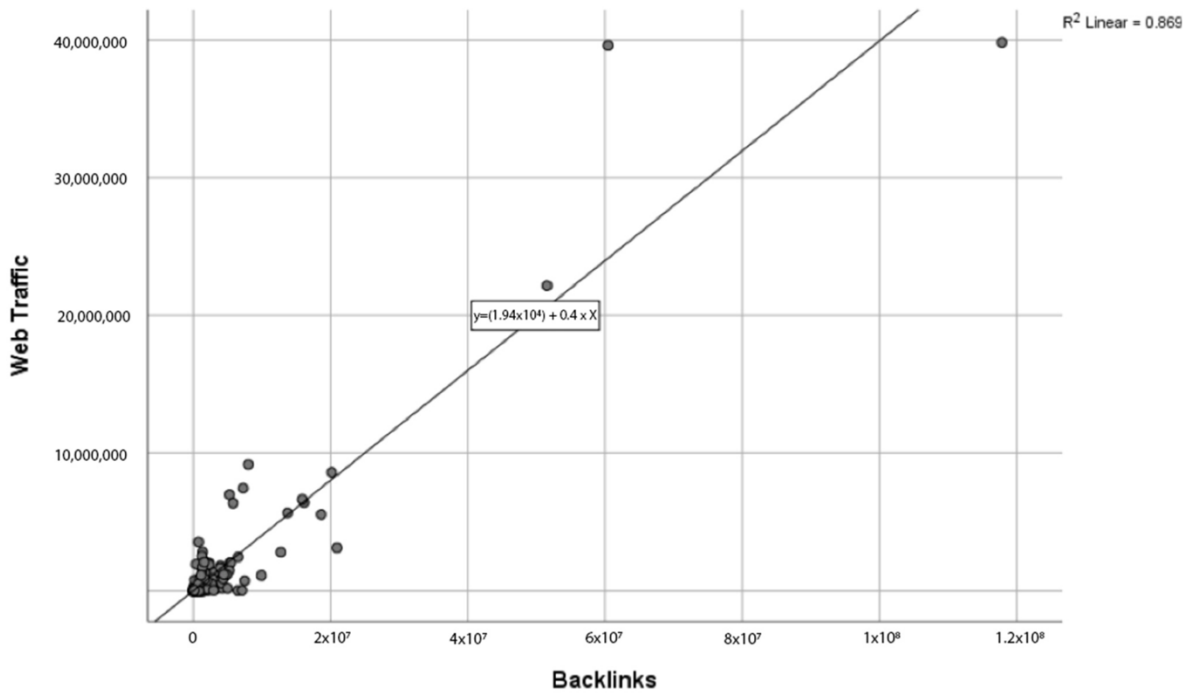
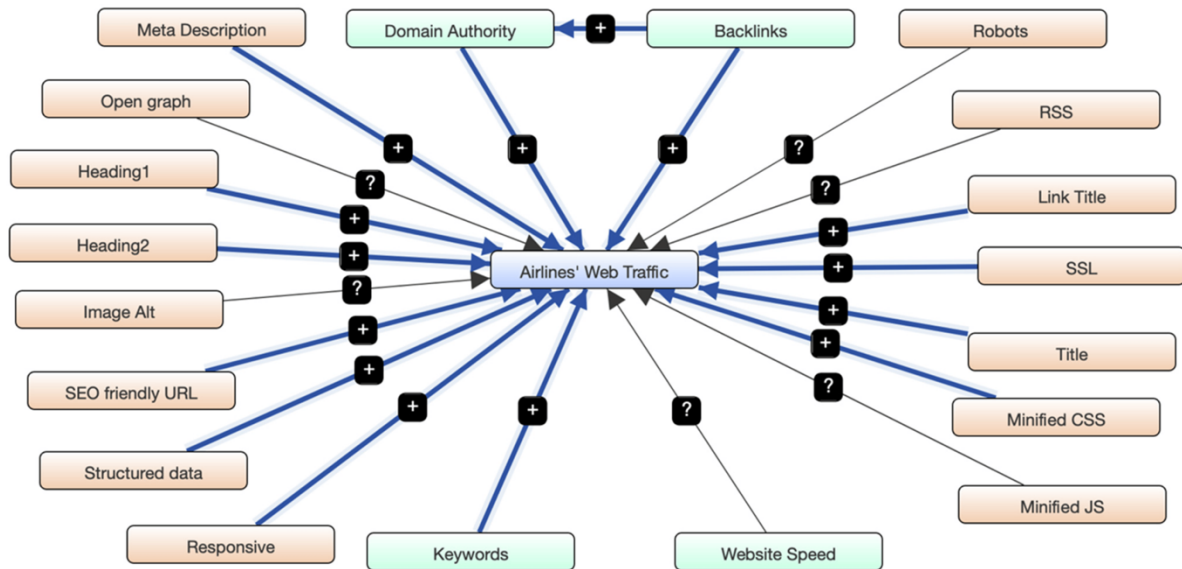


Figure 63. Scatterplot of Backlinks by web traffic.

### Diagnostic Exploratory Model Development

Results of the t-tests on SEO techniques and ANOVA, coefficients, and scatterplots on SEO metrics present significant relationships with the Web traffic. Thus, we aim to provide additional insights to marketing strategists over the impact of SEO techniques and metrics usage on airline websites, in favor of their digital marketing strategy. Therefore, we deployed fuzzy cognitive mapping (FCM), which will implement selected web metrics from both SEO techniques and metrics depicting the relationships to Web metrics, so as to obtain a more adaptive model for process assessment. Fuzzy cognitive mapping deploys a descriptive and consolidated stochastic classification methodology,

mainly used to represent the correlations between airlines' web metrics [407]. We use the orange color to mark the SEO techniques, green color to mark the SEO metrics, and blue color to mark the airlines' Web traffic. Blue arrows in **Figure 64** illustrate variables with positive relations with Web traffic and black arrows illustrate variables with neutral relations with Web traffic. Development of the FCM was conducted via the Mental Modeler cloud-based application [408].



**Figure 64.** Using fuzzy cognitive map to locate the correlations between SEO techniques/metrics and Web traffic.

Fuzzy cognitive maps are fuzzy graph structures that represent causal reasoning. Exploratory modelling implementation is vital when digital marketing agencies need to make a decision or a digital marketing plan [409].

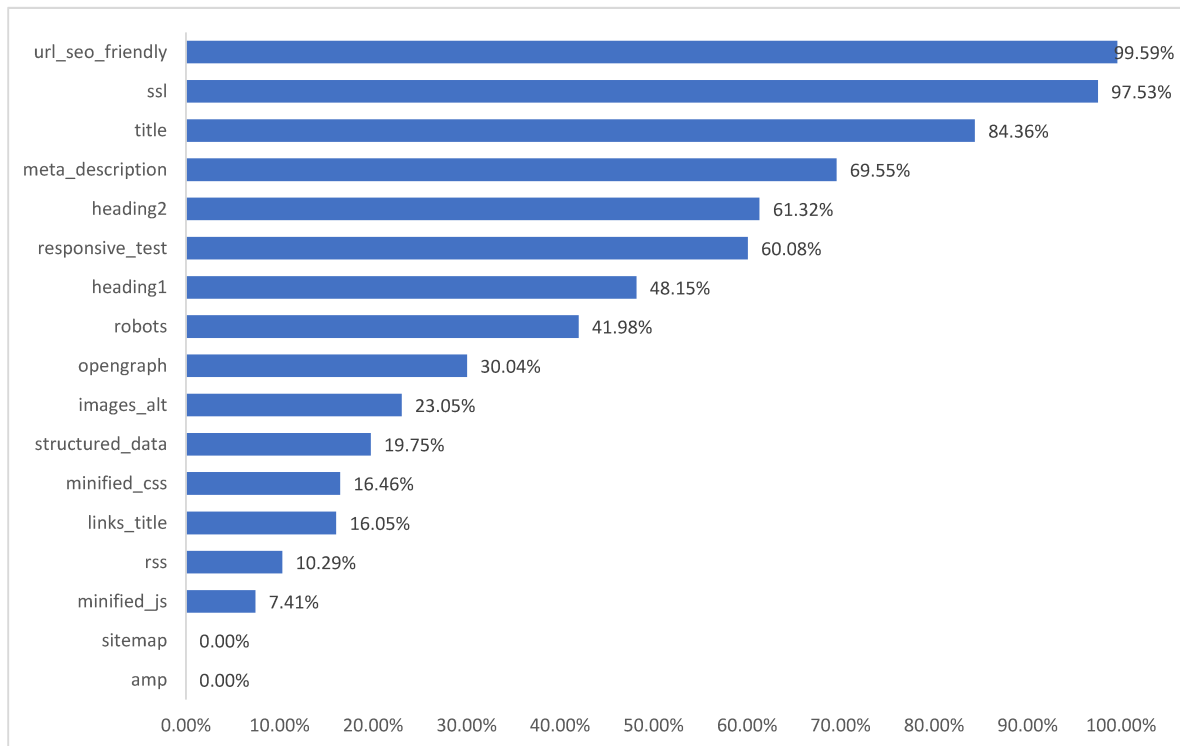
As we notice in **Figure 64**, most of the SEO techniques and metrics positively affect Web traffic. Some of them have a relationship with each other, such as Backlinks and DA.

#### 4.9.4. Discussion

The main focus of this study is the development of a precise methodology, containing pioneering context, aiming to provide useful insights concerning SEO usage and its contribution to the development of airlines' web traffic as well as airlines' sustainability. Using our own-developed tool, we retrieve airlines' data from IATA's website. For each airline website (243 in total), we performed SEO checks evaluating SEO techniques implemented on their source code. At the same time, we used four SEO third-party APIs to collect data for each airline's website, such as 12-month web traffic, DA, website's speed, and organic keywords. The average DA of the airlines' websites in our dataset is 56, the average loading speed is 10.5 s, and the average web traffic is 1,052,702 users per month.

We analyzed our data through descriptive analysis and t-tests to find out the contribution of each technique and metric to airlines' web traffic.

**Figure 65** presents the SEO techniques and airlines' website adoption scores. It was noticed that the majority of the airlines' websites implemented SEO techniques, such as SEO-friendly URLs by 99.59%, SSL certificates by 97.53%, title tag by 84.36%, and meta description by 69.55%. As mentioned earlier, the AMP and Sitemap SEO technique tests showed that none of the airlines' websites applied these techniques, although they are suggested by Google on Webmaster Guidelines [271].



**Figure 65.** SEO techniques adoption percentage.

Conducting descriptive regression analysis and t-tests accordingly in **Section 4.9.3**, we confirm the research hypotheses listed in **Problem Formulation and Research Hypotheses** section. Our study underlines that on-page SEO impacts airlines' website traffic on a great scale. The same applies to off-page SEO, where we confirm that backlinks influence airlines' web traffic. More specifically, when the beta value of backlinks increases by one unit, the web traffic will be increased by 0.399 units. From the organic keywords point of view, when the beta value of keywords increases by one unit, the airlines' web traffic will be increased by 10.24 units, respectively. On the contrary, website speed seems not to have any impact on airlines' web traffic, although mentioned as an important SEO factor in **Section 4.4.2**. In addition, through our analysis, it was found that there is a strong connection between backlinks and DA.

Although each individual SEO technique cannot influence the rankings on its own, the appropriate combination of SEO techniques can lead to greater search results. Expanding this research, future research could focus on how the combination of specific SEO techniques, creating potential groups, can affect the corresponding ranking in search results.

#### 4.9.5. Section Conclusion

COVID-19-era travel restrictions have led airline firms to adopt new digital marketing strategies to survive against their rising competition. Most airlines invested in SEO to increase their websites' rankings on SERPs and, by extension, to improve online visibility, attract more visitors, and increase bookings through their service-oriented e-commerce platforms. This study undertakes to uncover the SEO strategies airline websites follow. Using our own-developed tool, we scan each website's source code, mining the corresponding SEO techniques used by airlines. At the same time, we use four third-party APIs to collect valuable time-accurate data, such as DA, organic keywords, backlinks, and 12-month period web traffic. Meanwhile, through our data analysis, we end up with the most-adopted SEO techniques and how each one of them can affect airlines' websites' traffic. We conclude that SEO techniques and metrics have a great correlation with web traffic, which, in turn, increase airlines' conversions and bookings. Our research identifies the SEO techniques that are most

widely used, assuming that to be applied by large companies, such as airlines, they can also deliver results on smaller websites.

#### **4.10. *Greek Hotels' Web Traffic: A Comparative Study Based on Search Engine Optimization Techniques and Technologies***

##### **4.10.1. Section Summary and Significance of the Study**

Currently, websites rely heavily on digital marketing, notably SEO, for success. In the COVID-19 era, hotels have to employ every feasible means to stay afloat despite the bleak business and travel conditions. Many of them have already invested in digital marketing, especially SEO, by applying SEO techniques to their websites to attract more visitors and bookings through their service-oriented e-commerce platforms. Following our research in this section, we focus on examining the hotels' websites regarding the SEO techniques they have applied and their impact on web traffic to their websites. During a one-year observation period (February 2021–February 2022), we collected and analyzed web data from 309 top-listed Greek hotels using our own-developed software [17]. By creating and following a specific methodology, we came to valuable conclusions. In addition, we used fuzzy cognitive mapping to develop an exploratory model. From the descriptive analysis and technical SEO perspective, we have concluded that hotels websites' traffic and, by extension, their long-term viability are inextricably intertwined. Existing and future SEO marketers may benefit from our research's time-accurate insights on hotel SEO tactics.

##### **4.10.2. Materials and Methods**

The purpose of this study is to propose an innovative methodology for implementing an efficient framework for comprehending the strong correlation between SEO and web traffic, thereby providing valuable insights for hotels' websites. We followed seven research steps to achieve the desired outcomes.

- i. **Problem Formulation and Research Hypotheses.** In this stage, we clearly define the necessity of earning more organic traffic using SEO techniques along with five hypotheses that must be confirmed from our descriptive analysis.
- ii. **Data Retrieval.** In this stage, we describe the software developed for the purpose of this study, and the external APIs used to collect additional information regarding our dataset.
- iii. **Data Normalization and Limitations.** In this stage, we divide our dataset into three groups (Metrics, SEO Techniques, and Web Traffic) facilitating the descriptive analysis that follows. Each group is analyzed using different procedures depending on its type. Metrics and SEO Techniques were analyzed against the dependent variable Web Traffic.
- iv. **Descriptive Statistics.** Using the SPSS software, we perform descriptive analysis on our dataset.
- v. **Inferential Statistics**
  - o **T-Tests on SEO Techniques.** In this stage, we perform independent t-tests on each SEO technique by using the SPSS software.
  - o **ANOVA, Coefficients and Scatterplots on SEO Metrics.** In this stage, independent ANOVA, Coefficients and Scatterplots performed on each SEO Metric by using the SPSS software.
- vi. **Understand cause-and-effect relationship of SEO Techniques and Web Traffic.** In this stage, we conclude that the majority of SEO Techniques positively affect web traffic.
- vii. **Understand cause-and-effect relationship of SEO Metrics and Web Traffic.** In this stage, we conclude that the majority of SEO Metrics positively affect web traffic.

### Websites' Traffic Sources and Key Performance Indicators

Hotels' online presence and bookings are directly affected by various factors that should be analyzed and factored into our estimation. To obtain a more detailed picture of the hotels' online presence, we use their organic web traffic sources, which indicate the primary traffic source to their websites. As mentioned in **Section 4.1**, organic traffic is the traffic generated by users who use search engines (e.g., Google, Yahoo, etc.). It is an unpaid form of web traffic that plays a significant role in business success. Most websites rely on organic traffic results, which account for more than 60% of total website traffic [380]. As an unpaid source of traffic, organic traffic offers numerous benefits to businesses, including a lower initial investment than PPC campaigns, long-lasting results that can be enhanced year after year with the appropriate effort, and a significant increase in brand awareness [381].

Organic keywords, backlinks, website speed, and DA are four additional metrics that search engines consider when evaluating a website before ranking it in the SERPs [382]. Organic keywords are those for which websites are ranked in search engines' organic SERPs [384]. Backlinks are links from third-party websites to the target page that endorse that page [274]. Website speed is the time it takes for a web page to fully download and load from the webserver to the user's browser and it is measured in seconds. DA is a Moz-developed search engine ranking score that forecasts a website's strength and its ability to rank in SERPs. The DA score is a numeric value between 1 and 100, with higher values indicating a greater likelihood of ranking [367]. DA is not a metric that search engines can see. It attempts to mimic how search engines rank web pages. In the majority of cases, the DA score is indicative. The DA is critical when comparing it to other websites. Websites with a higher DA are more likely to appear higher in SERPs than those with a lower DA.

**Table 32** summarizes the web traffic metrics proposed as key performance indicators (KPIs) in this study.

**Table 32.** Description of the examined web analytics metrics and suggested KPIs and Performance Measurements.

KPIs (Unit)	Description
<b>Organic Traffic/Month (number of organic visitors per month)</b>	Organic search refers to the non-paid search results from a search engine. These results cannot be bought or influenced by advertisers, so they are the ones the search engine considers most relevant to the user's search query [386]. Organic traffic is described by the number of unique visitors per unit of time resulting from users' queries on search engines.
<b>DA (score in the range [0-100], integer)</b>	DA score can be used when comparing websites or tracking the "ranking strength" of a website over time [20,367]. Websites with higher DA are more likely to rank higher in SERPs than those with lower DA. Search engines do not use the DA metric to rank web pages to the search results. DA is a simulation metric made by Moz's SEO experts to model search engines' algorithms behavior [20,293].
<b>Website Speed (seconds)</b>	Website speed refers to how quickly a browser can load fully functional web pages from a given website.
<b>Organic Keywords (number of keywords)</b>	Organic keywords are keywords used in SEO to attract "free" traffic. When a user types a keyword on the search engines, it returns as a result websites which rank for this specific keyword [383]. SEO experts are constantly trying to detect high-traffic keywords that users prefer to search on search engines. The keywords are selected based on the website's niche. By creating landing pages, and using copywriting and on-page SEO techniques, SEO experts attempt to rank the website to specific keyword queries on search results. The more keywords rank for, the higher organic traffic received. SEO software tools, such as Ubersuggest [391], are used by SEO experts to monitor in which exactly

	keywords the website is ranking for and in which specific ranking position.
<b>Backlinks (number of backlinks)</b>	Backlinks are links that the considered page receives from other page(s). According to Google's founders, anchored links are links to a website containing the target keyword of the destination website. Anchors often provide more accurate descriptions of Web pages than the pages themselves [410]. Backlinks help secure a higher SERP ranking [327]. Backlinks affect mostly the position of a website followed by relevant content [387].

A performance indicator, also known as a key performance indicator (KPI), is a type of performance measurement that assesses a business's success in specific activities that it manages [384,385]. **Table 32** contains the KPIs examined in this study in addition to the web metrics. Monthly traffic analysis is beneficial for exporting more accurate results about a website's traffic and ranking performance.

### Sample Selection, Data Retrieval

We obtained data for this study from Google Maps and the Google Maps API [411]. The sample consists of 309 Greek hotels' websites which have user ratings greater than three out of five stars. For the above hotels' firms, data were collected concerning their websites' organic traffic, website speed, number of organic keywords, reviews, and their DA. Sample considered as representative providing with knowledge about the SEO applied by hotels as well as how this translated into traffic. Monthly data collection from hotels websites allowed for a more precise examination and comprehension of web metrics variance. The testing period has been extended to 12 months, with specific dates between February 2021 and February 2022. Because this period corresponds to the COVID-19 transportation restrictions, the results will provide a clear picture of the decline in bookings during this period.

We extended the PHP-based web crawler tool created in **Sample Selection and Data Retrieval** section to collect samples that extracted information about each hotel listed on the Google Maps website, including its legal name and website [16,20]. Following data extraction, we stored the entire dataset in a MySQL database, later used to perform SEO checks [412].

The software tool created for the purpose of this study draws the dataset from Google Maps API and stores each result in a MySQL database. It then traverses the dataset and, using cURL, extracts the source code from each hotel's website. The cURL project is computer software that includes both a library (libcurl) and a command-line tool (curl) for data transfer over a variety of network protocols [390]. Each of the fifteen SEO techniques discussed in **Section 4.2** has its function that searches the source code of the targeted website to determine whether the SEO Technique has been applied. Finally, the database is updated with the results of the checks. **Figure 66** illustrates the generated code and its primary functions in detail.

```
require ('seo-techniques-functions.php'); // ToDo! Get SEO techniques' functions
$return = retrieve_data_from_google_maps(); /* ToDo! Retrieve data from Google
Maps API */
$websites = create_dataset_array($return); /* ToDo! Create an array with our da-
taset */
foreach($websites AS $id => $website) { // ToDo! Run through our dataset
    $curl_results = do_curl_website($website); // ToDo! Do curl on current website
    $seo_checks = do_seo_checks($curl_results); /* ToDo! Do SEO checks using the
getElementByTagName */
```

**Figure 66.** PHP-based SEO software.

To supplement our research and to validate our findings, we used four third-party API tools: the Mobile-Friendly Test Tool API, the PageSpeed Insights API, the Mozscape API, and the Ubersuggest Traffic Analytic tool. Each of these tools obtains the URL of the website and returns results based on its measurements.

- The Google-created Mobile-Friendly Test Tool API validates a URL against responsive techniques. It returns a list of any mobile usability issues that may affect a user visiting the page on a mobile device [338].
- Google's PageSpeed Insights API analyses the performance of a web page and provides recommendations on how to improve various aspects of the page's performance, including page speed, accessibility, and SEO [378].
- The Mozscape API, developed by MOZ, accepts a website's URL as an input and returns accurate metrics such as DA [367].
- Ubersuggest, developed by Neil Patel, retrieves and returns a 360-degree view of any website's metrics and its sources [391].

After completing SEO audits on 309 hotels' websites, the exported data was imported into IBM SPSS 27. IBM SPSS 27 is one of the most widely used statistical software for ad hoc analysis, hypothesis testing, and predictive analytics. IBM SPSS Statistics is used by organizations to comprehend data, analyze trends, forecast, and plan to validate assumptions and reach accurate conclusions [392]. A descriptive analysis of various correlations confirmed the following research hypotheses. Finally, a fuzzy cognitive map (FCM) was created to visualize the relationship between SEO techniques, traffic, and backlinks, demonstrating the direct correlation between SEO techniques and traffic and, thus, with hotels' bookings and sustainability.

### **Problem Formulation and Research Hypotheses**

Increased competition, the COVID-19 era, and global travel restrictions tend to increase businesses' efforts to be more efficient in their digital marketing campaigns. Hotels make extensive use of every available tool, from copywriting to email marketing, social media marketing, and SEO, to reach out to potential customers in various ways. By understanding the consumer habits of their audience, hotel marketing agents develop marketing strategies to target this audience more effectively.

Hotels' websites are the primary source of bookings, and companies devote resources to organizing and optimizing the efficiency and effectiveness of their websites. Their position on the SERPs plays a crucial role in gaining organic visitors. The web page title, meta description, and even the rich results must all be applied with care to earn the coveted click. The content, user interface, and website speed work together to pique the visitor's interest and convert them into loyal customers. Given that hotel websites' primary objective is traffic and conversion to clients, we define five hypotheses and attempt to confirm them to expand practical knowledge about the significance and impact of SEO on hotel websites traffic.

- **Hypothesis 1 (H1).** SEO techniques affect the amount of traffic that hotels' websites receive.

The primary objective of this study is to determine whether SEO has an impact on the amount of traffic that hotels' websites receive. By deciding which SEO techniques hotels have implemented on their websites and comparing the results to the increase in traffic, we can gain valuable insight into the potential benefits of SEO techniques. As mentioned in **Sample Selection, Data Retrieval** section, a dataset was selected from Google Maps consisting of 309 Greek hotels' websites, which have user ratings greater than three out of five stars. Scanning and performing SEO tests on each website source code, using our own developed software [412], we gather information regarding the SEO techniques each website has implemented. Using third-party software tools described in **Sample Selection, Data Retrieval** section, we scan each website, gathering information concerning their DA, website speed, organic keywords, organic traffic, and backlinks. All the data collected is stored in an excel sheet. Using the IBM SPSS 27 software, a descriptive analysis was conducted in order to find connections/patterns

between SEO techniques used by websites and their organic traffic. The ultimate goal of this research is to uncover which SEO techniques used by Greek hotels have a more positive effect on organic traffic, assuming that the websites with the highest organic traffic have applied the appropriate SEO techniques in their source code.

- **Hypothesis 2 (H2).** On-page SEO techniques affect the ranking of a hotel’s website.

Hypothesis 2 is a segment of Hypothesis 1, targeting the on-page SEO techniques. Our second hypothesis is based on the effects of on-page SEO techniques on website ranking. Multiple keyword targeting and on-page SEO techniques are assumed to increase organic results.

- **Hypothesis 3 (H3).** Off-Page SEO techniques—backlinks affect the ranking of a hotel’s website.

Hypothesis 3 is a segment of Hypothesis 1, targeting the off-page SEO techniques. Hypothesis 3 examines off-page SEO techniques and their effect on the traffic to hotel websites. We assume that creating backlinks to third-party websites will boost a website’s DA, which will result in higher rankings in the SERPs.

- **Hypothesis 4 (H4).** Hotels’ organic keywords affect their website’s ranking.

We assume in Hypothesis 4 that organic keywords are valuable for the hotels’ website, resulting in higher SERP rankings and increased web traffic.

- **Hypothesis 5 (H5).** Hotels’ backlinks affect their DA.

We assume in Hypothesis 5 that backlinks are intrinsically linked to DA.

#### 4.10.3. Results

##### Data Normalization

In this section we examine the collected data. As mentioned in previous sections, our tool performed SEO checks on 309 websites. The results were stored in excel so that we can use them in further data analysis [413].

Collected data are separated into three Groups.

- i. Group 1: Metrics are checks performed by third-party tools and are not SEO techniques. The contents of the Metrics Group are as follows: DA (integer 0 to 100), speed\_test (decimal), number of keywords (integer), and number of backlinks (integer).
- ii. Group 2: SEO Techniques are checks performed by our tool regarding the existence (or not) of SEO techniques on each website of the dataset. All metrics returned 1 if the SEO technique has been applied to the website, 0 otherwise. The contents of the SEO Techniques Group are as follows: images\_alt, links\_title, rss, sitemap, robots, heading1, heading2, web\_ssl, meta\_description, opengraph, url\_seo\_friendly, amp, minified\_css, minified\_js, title, structured\_data, and responsive\_test.
  - a. To extend the results of our research, by giving readers additional insights related to SEO, we introduce two variables which are not SEO techniques but can be treated statistically in the same way. The first variable is the well-known open-source CMS WordPress. We present this variable to identify if WordPress CMS can provide better results regarding SEO and organic traffic than other web platforms. The second variable is the Yoast SEO plugin. Yoast SEO plugin is a WordPress extension that uses advanced suggestion tools to help website’s administrator improve website’s content and structure proposing changes such as keyword targeting, internal linking, structured data, sitemap, etc. We present this variable to determine if the Yoast SEO plugin can deliver better results regarding SEO and organic traffic if installed in a WordPress CMS. In Group 2, two more parameters are added, «is\_yoast» and «is\_wp», with a value of 1 for those websites that use them and a value of 0 for those websites that do not use them.
- iii. Group 3: Web Traffic (organic traffic) are the monthly statistical data collected for each website. All web traffic data are integers and refer to websites’ data for the last twelve months. The

contents of the Web Traffic Group are as follows: traffic\_1, traffic\_2, traffic\_3, traffic\_4, traffic\_5, traffic\_6, traffic\_7, traffic\_8, traffic\_9, traffic\_10, traffic\_11, and traffic\_12.

### Limitations

During the examination of the responsive\_test technique, it was found that the Mobile-Friendly Test Tool API created by Google considers non-responsive those pages that have even one non-responsive element. Manual checks revealed that these specific websites partially adopt responsive techniques and have several elements in their source code that are non-responsive. These pages are marked with the number 0 which means that they do not follow that specific SEO technique. Additionally, in 22 out of the 309 websites, the speed measurements were performed manually with the tool Pingdom website speed test by Solar Winds, because the PageSpeed Insights API by Google could not operate speed test on these pages.

During the AMP and Sitemap SEO technique tests found that none of the websites have been applied these techniques, although they are suggested by Google on Webmaster Guidelines [271]. Consequently, t-tests cannot be performed on these techniques excluding them from our final results. We collected the monthly organic traffic for 12 months to have a more representative view of the website traffic. For the data analysis that follows, we calculated the mean value of 12 months of organic traffic for each website and saved it in a new column called Web Traffic. All the following comparisons regarding web traffic refer to the mean value of organic traffic for each website.

Raw data collected from multiple websites do not convey any significant trends or the behavior of the individuals before the corresponding analysis. Data analysis refers to the process of converting the raw data into meaningful information, using the mathematical, statistical, or computational algorithms for better comprehension [394]. For the quantitative data analysis, some major steps need to be considered to firmly execute the assessment and generate the intended results. The procedure consists of two major steps: (1) reviewing the data by descriptive analysis and (2) conducting inferential and descriptive statistics to answer the research questions [395]. The study is presented in the following sections of the chapter and each section is intended to answer the research questions as the analysis must be aligned with the research objectives.

### Descriptive Statistics

Descriptive analysis helps to describe, demonstrate or summarize the collected data in a constructive manner so that the trends and patterns can be easily observed and analyzed [396]. Mean, median, mode, standard deviation, skewness, and kurtosis are some of the important measures. Mean, median and mode measure the central tendency of the variable to typify the whole set of data [414]. Similarly, skewness is the measure of the degree of lopsidedness in the distribution on the frequency, and kurtosis is the measure of the degree of tailedness in the frequency distribution [415]. Standard deviation measures the risks, volatility, scatteredness, or variability in the data [416]. **Table 33** indicates that the maximum web traffic, DA, speed test, keywords and backlinks of this study are 1,058,500, 90, 15.4, 2,898,545 and 32,965,989, respectively. The values of skewness or kurtosis within -1.5 and +1.5 are considered right skewed distribution of the data [397].

**Table 33.** Descriptive statistics on SEO metrics.

	N	Minimum		Maximum		Mean		Std. Deviation		Skewness		Kurtosis	
		Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error		
Web Traffic	309	2	1,058,500.00	7939.3091	69,130.01356	13.320	0.139	188.769	0.276				
DA	309	3	90	28.81	11.472	0.907	0.139	4.153	0.276				
Website Speed	309	0.4	15.4	3.501	2.5462	1.781	0.139	3.589	0.276				
Organic Keywords	309	0	2,898,545	27009.03	263236.714	10.153	0.139	102.411	0.276				
Backlinks	309	3	32,965,989.00	218,197.9773	2,315,388.35	12.246	0.139	157.348	0.276				

## Inferential Statistics

Inferential statistics refers to the procedure of using data analysis to infer attributes of an underlying probability distribution. Inferential statistics deal with the information acquired from the sample of the population to draw conclusive statements about the entire population [399]. Inferential statistics make use of statistical models to compare the sample data to that of earlier studies. It varies from descriptive statistics in that it allows one to draw conclusions based on extrapolations rather than merely reporting the data that has been seen, like descriptive statistics does [399]. From our hypotheses, we assume that SEO Metrics and SEO Techniques are positively affecting the dependent variable Web Traffic. Using inferential statistics, we compare each SEO metric and technique against the web traffic and—without loss of generality—we extend the corresponding conclusion to the larger population.

### *Impact of SEO Techniques Group on the Web Traffic*

At this point, 15 different t-tests were performed for each one of the 15 SEO techniques (Group 2) using the SPSS 27 software by IBM. After this, the t-tests mean values observed. The mean value is equal to the sum of all the values in the data set divided by the total number of values. The “YES” mean reflects the average web traffic from the websites that applied the technique. The “NO” mean reflects the average web traffic from the websites that have not applied the technique. If the “YES” mean is greater than “NO” mean the websites which have implemented this technique in their source code have greater traffic and the technique is valuable for the website. Another component to consider when comparing Yes or No is the p-value. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis [400]. A p-value less than 0.05 (typically  $\leq 0.05$ ) is statistically significant. A p-value higher than 0.05 ( $>0.05$ ) is not statistically significant and indicates strong evidence for the null hypothesis [401]. The results from 2 of 15 t-tests are presented in detail below, and the rest are presented briefly in **Table 36** for the sake of brevity.

**Table 34.** Independent t-test on the meta description technique.

Technique	Mean (Number of Visitors per Month)		Standard Deviation (Number of Visitors per Month)		t-Value	p-Value
	Yes (n = 193)	No (n = 116)	Yes (n = 193)	No (n = 116)		
Meta Description	8739.23	6608.39	77063.28	53677.81	-0.262	0.02

### **Impact of meta description technique on the web traffic**

- **H0.** There is no difference in websites traffic due to the meta description technique (null hypothesis).
- **H1.** There is a difference in websites traffic due to the meta description technique.

In this study, an independent samples t-test is applied to draw a comparison of the mean web traffic between websites that implement meta description technique (n = 193) and websites that did not implement the meta description technique (n = 116) (**Table 34**). The t-test was statistically significant, with the mean web traffic of meta description technique-implemented websites (M = 8739.23, SD = 77,063.28) being higher than the meta description technique unimplemented websites: (M = 6608.39, SD = 53,677.81),  $t = -0.260$ ,  $p < 0.05$ , two-tailed. Therefore, the null hypothesis, H0, which states that there is no difference in websites traffic due to the meta description technique, can be rejected. Consequently, it can be concluded that websites which have implemented the meta description technique have greater traffic, and by extension, this technique is valuable for the websites.

**Table 35.** Independent t-test on the Robots.txt technique.

Technique	Mean (Number of Visitors per Month)	Standard Deviation (Number of Visitors per Month)	t-Value	p-Value
-----------	-------------------------------------	---	---------	---------

	Yes (n = 196)	No (n = 131)	Yes (n = 196)	No (n = 131)		
Robots.txt	6094.49	11,139.16	42,905.73	99,602.74	0.617	0.04

### Impact of Robots.txt technique on the web traffic

- **H0.** There is no difference in websites traffic due to the Robots.txt technique (null hypothesis).
- **H1.** There is a difference in websites traffic due to the Robots.txt technique.

In this study, an independent samples t-test is applied to draw a comparison of the mean web traffic between websites that implement the Robots.txt technique (n = 196) and websites that did not implement the Robots.txt technique (n = 131). The t-test was statistically significant, with the mean web traffic of robot technique-implemented websites (M = 6094.49, SD = 42,905.73) being lower than the robot technique unimplemented websites: (M = 11,139.16, SD = 99,602.74),  $t = 0.617$ ,  $p < 0.05$ , two-tailed. Therefore, the null hypothesis that there is no difference in websites traffic due to the robot technique was rejected. Therefore, it can be concluded that websites that have implemented Robots.txt technique have failed to reach more web traffic and the null hypothesis H0 cannot be rejected.

**Table 35** presents the 15 t-tests performed for the 15 corresponding SEO techniques. In the last column of **Table 35**, the SEO techniques that resulted in providing a positive impact on the traffic are marked as 1, while the techniques that do not have a significant contribution to the traffic are marked as 0. At the end of **Table 36**, two more t-tests were performed regarding WordPress CMS and Yoast SEO against the dependent variable Web Traffic.

**Table 36.** T-test on 15 SEO Techniques.

Technique	Mean (Number of Visitors per Month)		Standard Deviation (Number of Visitors per Month)		t-Value	p-Value	Valuable
	Yes	No	Yes	No			
Meta Description	8739.23	6608.39	77,063.28	53,677.81	-0.262	0.02	1
Image Alt	16,867.69	1791.89	107,719.97	4978.16	-1.892	0.03	1
SEO friendly URL	7964.38	215.42	69,241.10	0	-0.112	0.01	1
Minified CSS	19,867.59	2315.97	121,203.18	7434.48	-2.094	0.00	1
Title	10,237.10	1883.92	81,027.66	6470.74	-0.948	0.00	1
Structured data	13,017.17	5092.62	101,275.80	41,495.26	-0.967	0.00	1
Responsive	11,117.37	1764.77	84,892.54	5268.46	-1.127	0.02	1
Heading1	9941.26	4105.36	84,441.15	16,683.35	-0.704	0.04	1
Heading2	9711.04	6061.27	84,777.06	47,439.21	-0.463	0.00	1
SSL	8324.65	386.56	70,855.70	563.44	-0.433	0.01	1
Open graph	9066.93	6680.37	83,100.83	49,322.96	-0.303	0.02	1
Link Title	11,925.11	6877.51	71,897.13	68,486.96	-0.522	0.00	1
RSS	29,721.21	2690.65	154,236.56	11,543.47	-2.74	0.00	1
Robots	6094.49	11,139.16	42,905.73	99,602.74	0.617	0.04	0
Minified JS	23,685.24	2251.34	132,989.81	7181.28	-2.425	0.00	1
Is WordPress	2697.31	13,568.29	12,955.12	98,506.26	1.383	0.168	0
Is Yoast	2762.92	9511.88	7731.00	78,792.80	0.725	0.469	0

### Impact of SEO Metrics Group on the Web Traffic

Regression is a statistical technique to formulate the model and analyze the relationship between the dependent and independent variables. It aims to study the degree of relationship between two or more variables. This is achieved with the help of hypothesis testing.

In this part of our research, three different tests, one-way analyses of variance (ANOVA), coefficients and scatterplots, were performed for each of the four Metrics (Group 1) using the SPSS 27 software by IBM. After the analysis, it is intended to determine the relationship between Metrics and Web Traffic.

The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups, or not. Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all

other independent variables are held constant. A simple scatterplot can be used to (a) determine whether a relationship is linear, (b) detect outliers and (c) graphically present a relationship between two continuous variables [402].

**The null and alternate hypotheses remain the same as the Techniques.**

- **H0:** [Metric] doesn't have positive impact on the web traffic.
- **H1:** [Metric] has a positive impact on the web traffic.

The model summary table reports the strength of the relationship between the model and the dependent variable (Table 37).

- DA: The regression model summary has been explained by 16.7% in variability of the web traffic,  $R^2 = 0.167$ , adjusted  $R^2 = 0.165$ .
- Website Speed: The regression model summary has been explained by only 0.1% in variability of the web traffic,  $R^2 = 0.001$ , adjusted  $R^2 = -0.002$ .
- Keywords: The regression model summary has been explained by 72.7% in variability of the web traffic,  $R^2 = 0.727$ , adjusted  $R^2 = 0.726$ .
- Backlinks: The regression model summary has been explained by 45.2% in the variability of the web traffic,  $R^2 = 0.452$ , adjusted  $R^2 = 0.450$ .

**Table 37.** Model Summary. Strength of the relationship between the model and the dependent variable.

Metric	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
DA	1	0.409	0.167	0.165	63,180.47283
Website Speed	1	0.037	0.001	-0.002	69,195.23812
Organic Keywords	1	0.853	0.727	0.726	36,191.48936
Backlinks	1	0.672	0.452	0.450	51,279.24795

The ANOVA table represents the overall significance of the model, which is determined by the web traffic (Table 38). The F ratio is the ratio of two mean square values. Degrees of freedom (DF) of error has been calculated by Total sample ( $n = 309$ )-2 = 307 and total DF has been calculated by Total sample ( $n = 309$ )-1 = 308.

- DA: The result shows  $F(1, 307) = 61.738$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this model is related to the dependent variable.
- Website Speed: The result shows  $F(1, 307) = 0.420$ ,  $p > 0.05$ , which suggests the model is not statistically significant. Therefore, the independent variable in this is not related to the dependent variable.
- Keywords: The result shows  $F(1, 307) = 816.754$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this is related to the dependent variable.
- Backlinks: The result shows  $F(1, 307) = 252.759$ ,  $p < 0.05$ , which suggests the model is highly significant. Therefore, the independent variable in this model is related to the dependent variable.

**Table 38.** ANOVA results. Overall significance of the model.

Model		Sum of Squares	df	Mean Square	F	p-Value
DA	Regression	246,445,253,450.713	1	246,445,253,450.713	61.738	0.000
	Residual	1,225,474,049,215.856	307	3,991,772,147.283		
	Total	1,471,919,302,666.569	308			
Website Speed	Regression	2,009,142,195.171	1	2,009,142,195.171	0.420	0.518
	Residual	1,469,910,160,471.398	307	4,787,980,978.734		
	Total	1,471,919,302,666.569	308			
Organic Keywords	Regression	1,069,803,364,626.086	1	1,069,803,364,626.086	816.754	0.000
	Residual	402,115,938,040.483	307	1,309,823,902.412		
	Total	1,471,919,302,666.569	308			
Backlinks	Regression	664,643,992,835.479	1	664,643,992,835.479	252.759	0.000

Residual	807,275,309,831.090	307	2,629,561,269.808
Total	1,471,919,302,666.569	308	

The coefficient table represents how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant [403] (Table 39). When interpreting the coefficient table, the B variable is important, representing the increase in the dependent variable as soon as the independent increases by one unit. On the other hand, the beta variable compares the strength of the effect of each individual independent variable to the dependent variable. The higher the value of the beta coefficient, the stronger the effect, based on absolute numbers [404].

- DA: When the beta value of DA is increased by one-unit, web traffic will have increased by 2465.645 visitors per month. Therefore, DA does not have a positive impact on the web traffic null hypothesis is rejected. It can be concluded that the DA scores have a significant positive effect on the web traffic.
- Website speed: Since the p value of the following table is greater than 0.05, the impact of speed is not statistically significant. Therefore, the null hypothesis that speed test does not have a positive impact on the web traffic is true and can't be rejected.
- Keywords: When the beta value of keywords is increased by one-unit, the web traffic increases by 0.224 visitors per month. Therefore, keywords do not have a positive impact on the web traffic, and thus, the null hypothesis is rejected. It can be concluded that the keywords have a significant positive effect on the web traffic.
- Backlinks: When the beta value of backlinks is increased by one-unit, the web traffic increases by 0.020 visitors per month. Therefore, backlinks do not have a positive impact on the web traffic, and thus, the null hypothesis is rejected. It can be concluded that the backlinks have a significant positive effect on the web traffic.

**Table 39.** Coefficient results. Results show how much dependent variable is expected to increase when that independent variable increases by one.

Model	B	Std. Error	Beta	t	p-Value
(Constant)	63,093.600	9728.574		6.485	0.000
DA	2465.645	313.800	0.409	7.857	0.000
(Constant)	4427.210	6700.015		0.661	0.509
Website Speed	1003.086	1548.492	0.037	0.648	0.518
(Constant)	1892.320	2069.706		0.914	0.361
Organic Keywords	0.224	0.008	0.853	28.579	0.000
(Constant)	3561.608	2930.141		1.216	0.225
Backlinks	0.020	0.001	0.672	15.898	0.000

Scatterplots are essential for identifying trends and patterns. In a scatterplot, each observation (or point) has two coordinates [405]. The strength of the link between the variables is determined by calculating the correlation coefficient. The plot shows the first variable's value on the X axis and the second variable's value on the Y axis for each data point [406]. The relationship between two quantitative variables is shown in a scatter plot (Figure 67).

- DA: The following graph indicates the coordinates of DA and web traffic, since the web traffic has been gradually increased because of the increase in DA scores, a indicating linear relationship between DA and the web traffic.
- Keywords: The following graph indicates that the more keywords that have been used in the website, the more traffic it will get. Therefore, there is a strong positive linear relationship between keywords and web traffic.
- Website speed: The following shows the relationship between speed test and web traffic. It shows that the increase in speed test does not result in a proportional increase or decrease in

the web traffic. Therefore, there is a constant and nonlinear relationship between these two variables.

- Backlinks: Like keywords, the following graph indicates the more backlinks that have been used in the website, the more traffic it will get. Therefore, it can be concluded that there is a strong positive linear relationship between keywords and web traffic.

We used exactly the same data analysis methodology as the SEO metrics to identify the correlation between backlinks and DA. The findings of the regression model have found that backlinks have a strong connection with DA. When the beta value of backlinks is increased by one unit, on average, the DA will increase by 2.214 units.

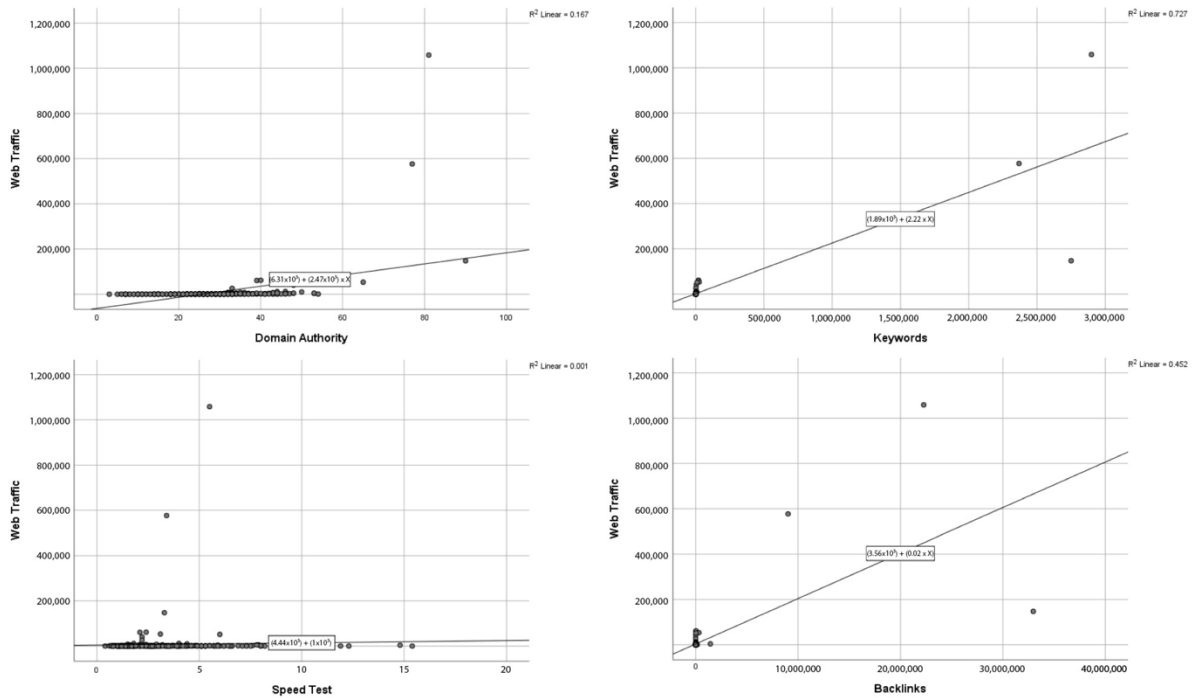
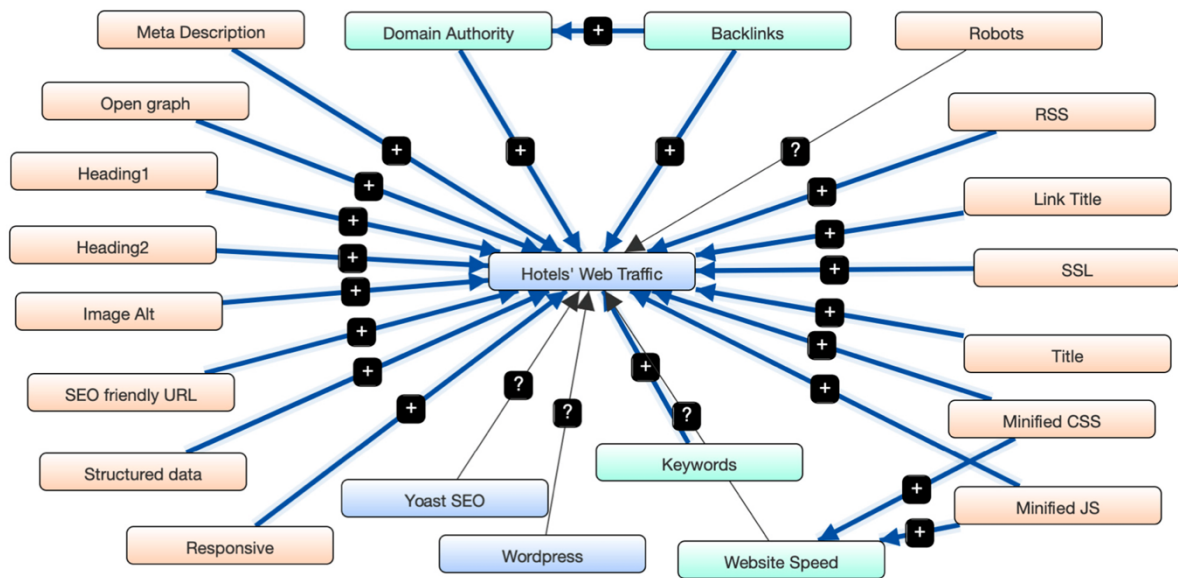


Figure 67. Scatterplot of each Metric by web traffic.

### Diagnostic Exploratory Model Development

Results of the t-tests on SEO Techniques and of the ANOVA, coefficients and scatterplots on SEO Metrics present significant relationships with the Web Traffic. Thus, we aim to provide additional insights to marketing strategists over the impact of SEO Techniques and Metrics usage on hotels' websites, in favor of their digital marketing strategy. Therefore, we deployed fuzzy cognitive mapping (FCM), which will implement selected web metrics of both SEO Techniques and Metrics depicting the relationships to Web Metrics, so as to obtain a more adaptive model for process assessment. Fuzzy cognitive mapping deploys a descriptive and consolidated stochastic classification methodology, mainly used to represent the correlations between hotels' web metrics [407]. We use the orange color to mark the SEO Techniques, green color to mark the SEO Metrics, and blue color to mark the hotels' Web Traffic. Blue arrows in **Figure 68** illustrate variables with positive relations with Web Traffic, and black arrows illustrate variables with neutral relations with Web Traffic. Development of the FCM has been conducted via the Mental Modeler cloud-based application [408].



**Figure 68.** Using fuzzy cognitive map to locate the correlations between SEO Techniques/Metrics and Web Traffic.

Fuzzy cognitive maps are fuzzy graph structures that represent causal reasoning. Exploratory modelling implementation is vital when digital marketing agencies need to make a decision or a digital marketing plan [409].

As we notice in **Figure 68**, most of the SEO Techniques and Metrics positively affect Web Traffic. Some of them have a relationship with each other, such as Backlinks and DA.

#### 4.10.4. Discussion

The primary objective of this study is to develop a precise methodology grounded in a pioneering context, to provide actionable insights into the use of SEO and its contribution to the growth of hotels' web traffic and sustainability. We retrieve hotel data from Google Maps API using a tool we developed. We performed SEO checks on each hotel's website (309 in total), evaluating the SEO techniques implemented in their source code. We also collected data for each hotel's website using four SEO third-party APIs, including 12-month web traffic, DA, website speed, and organic keywords. The average DA of our dataset's hotels' websites is 28, the average page load time is 3.5 s, and the average monthly web traffic is 7939 users.

We used descriptive analysis and t-tests to determine the relative contribution of each technique and metric to hotel web traffic.

**Figure 69** illustrates the adoption rates of SEO techniques and hotel websites. The majority of hotels' websites implemented SEO techniques such as SEO-friendly URLs (99.68%), SSL certificates (95.15%), title tag (72.49%), and Heading 1 (65.70%). As previously stated, the AMP and Sitemap SEO technique tests revealed that none of the hotels' websites utilized these techniques, even though Google recommends them in its Webmaster Guidelines [271]. The descriptive analysis also uncovered that hotel websites that are based on the open-source CMS WordPress (51.78% of our dataset) do not present more organic traffic than other websites. Additionally, the websites that use the Yoast SEO plugin (23.30% of our dataset) do not show more organic traffic than the other websites, which is in contrast to the so-called plugin that promotes SEO.

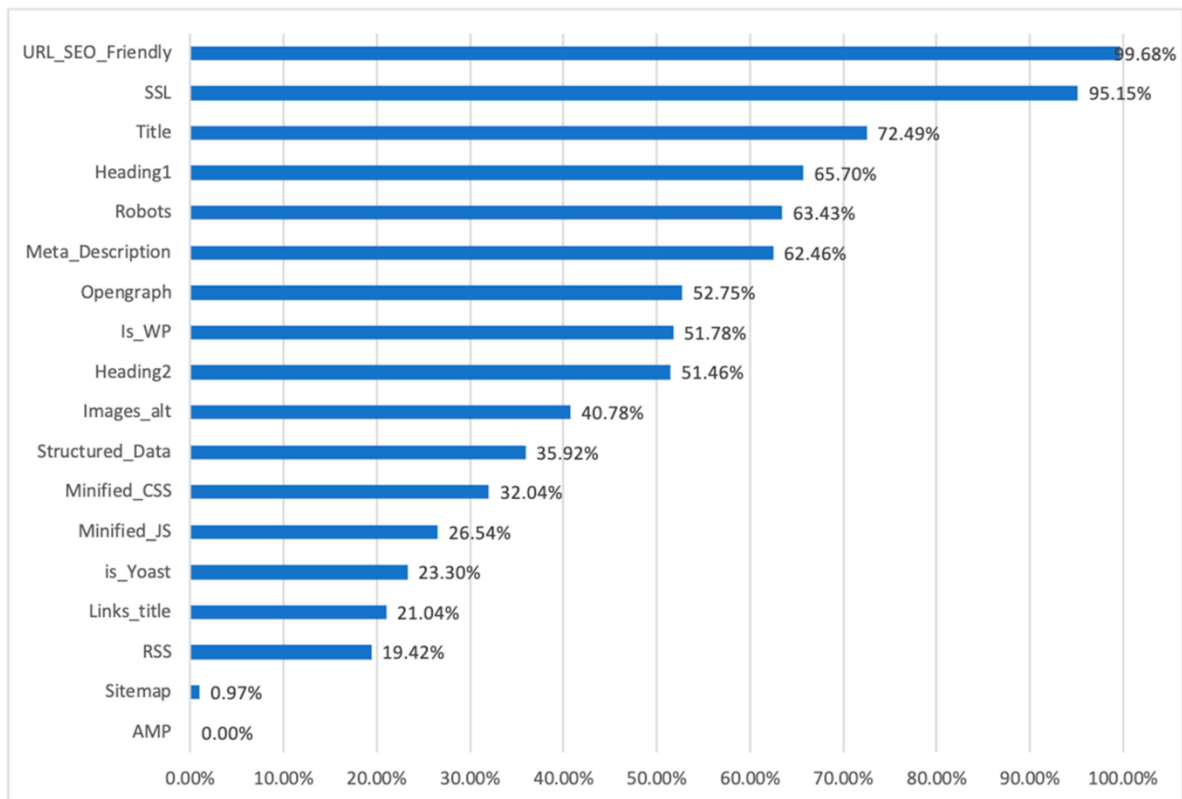


Figure 69. SEO Techniques adoption percentage.

Section 4.10.3 confirms the research hypotheses listed in **Problem Formulation and Research Hypotheses** section by conducting descriptive, regression, and t-test analysis. Our study demonstrates that on-page SEO has a sizable impact on the hotels' website traffic. The same is true for off-page SEO, where we confirm that backlinks affect hotels' websites. More precisely, increasing the beta value of backlinks by one unit results in an increase in web traffic of 0.020 visitors per month. When the beta value of keywords is increased by one unit, the hotels' web traffic increases by 0.224 visitors per month. On the contrary, despite being mentioned as a critical SEO factor in **Section 4.4.2**, website speed appears not to affect hotels' web traffic. Additionally, our analysis revealed a strong correlation between backlinks and DA.

#### 4.10.5. Section Conclusion

Travel restrictions and decline during the COVID-19 era have compelled hotel companies to develop new digital marketing strategies to compete with rising competition. The majority of hotels have invested in SEO to improve their online visibility, attract more visitors, and increase bookings through their service-oriented e-commerce platforms.

This study aims to ascertain the SEO strategies used by hotels' websites. We scan each website's source code, mining the corresponding SEO techniques used by hotels, using our own-developed tool. Simultaneously, we collect valuable time-accurate data from four third-party APIs, including DA, organic keywords, backlinks, and 12-month web traffic. Meanwhile, through our data analysis, we determine the most widely used SEO techniques and the impact on the traffic to hotels' websites. We conclude that SEO techniques and metrics strongly correlate with web traffic, which results in increased conversions and bookings for hotels.

## 4.11. The Search Engine Optimization Story of an old-fashioned E-commerce

### 4.11.1. Section Summary and Significance of the Study

As mentioned in previous sections, SEO encompasses a set of techniques and principles employed on websites to enhance their search performance and rankings. [417]. The SEO strategy aims to achieve a higher ranking for the website in specific keywords and to boost its organic visitors [418]. **Section 4.2** and **Section 4.3** present some of the most crucial SEO techniques. In this section, we apply these previously discussed SEO techniques to a live E-commerce to assess and validate their impact on organic traffic. Following the implementation of these SEO techniques, organic traffic increased by 782%, and the E-commerce currently ranks on the first page of search results for multiple keywords.

### 4.11.2. Materials and Methods

To demonstrate the effectiveness of the aforementioned techniques and technologies, SEO was applied to a live E-commerce. This E-commerce (palaio.gr) and the associated company actively engage in the sale and restoration of old objects. The primary objective of the E-commerce platform was to boost sales by showcasing its products on the Internet.

The legacy website, created in 2011 using a content management system (CMS) from that era, faced challenges despite the webmaster's efforts to increase website traffic. Daily uploads of new products and content did not yield the desired success. Even after multiple AdWords campaigns, the website only managed to reach the 4th page, specifically position 37, in Google searches for the keyword "Παλαιοπωλείο" (translated to "Antique Store" in English). Additionally, many of the website's products were not even indexed by search engines, making them difficult to find using keywords other than the keyword "Antique Store". In 2011, users did not frequently use mobile devices to search the Internet, which did not negatively impact the website's search engine ranking until 2016. However, with more than 25% of users browsing the Internet via mobile devices since 2016, traffic and sales fell dramatically, due to the absence of a mobile-friendly version of the website. By the end of 2018, the website received a maximum of 118 people per week, as illustrated in **Figure 70**. The business owner's goal was to attract more organic visitors and convert them into internet sales. Ensuring that visitors could browse the website seamlessly on mobile devices, smartphones, and tablets became imperative. Furthermore, the website needed to load faster and achieve rankings not only for the keyword "Antique Store", but also for related contextual keywords.

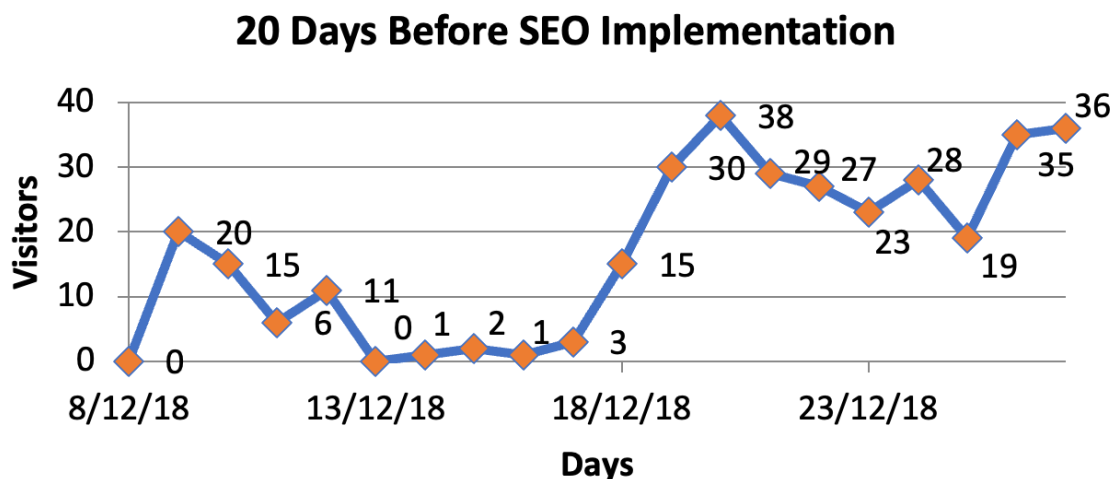


Figure 70. Organic Traffic per Day Before SEO Implementation - Statistics Exported from Google Analytics

### 4.11.3. Search Engine Optimization Results

At the beginning of 2019, a new E-commerce website for the business was launched on the same domain name. The newly developed website intentionally did not use any CMS platform. It was

crafted using responsive web technologies, incorporating custom HTML/CSS, JavaScript/jQuery, PHP, and MySQL to ensure a lighter and more user-friendly experience. Despite being created from scratch, featuring enhanced mobile responsiveness and a modern appearance, we deliberately retained the previous appearance, to avoid surprising returning visitors. For instance, we did not alter the positions of items such as menus and the search bar, striving to design the home page to closely resemble the old one. Simultaneously, throughout the development phase, we implemented all the SEO techniques and technologies outlined in **Sections 4.2, 4.3, and 4.4.**

### **Responsive Design - Mobile-Friendly Design**

According to the statistics gathered by Google Analytics for the website, it is evident that during the last week of 2018, 55% of users visited the website from desktop or laptop, 34% from smartphones, and 11% from tablets. As of July 2019, 42% of users access the website from desktop or laptop, 49% from smartphones, and 9% from tablets. The implementation of responsive design on the website has likely contributed to an increase in both traffic and session duration.

### **Keyword Ranking**

At the beginning of 2019, the website was positioned in the middle of the 4th page in searches using the keyword “Antique Store” (actually for the corresponding Greek term, i.e., “Παλαιοπωλείο”). Presently, the website holds the first position in searches on the first page of Google. Through keyword analysis and comprehensive SEO applied to the website, it has achieved the top position in searches for numerous Greek keywords such as “old jar” (i.e., “παλαιό πιθάρι” in Greek), “old furniture” (i.e., “παλαιά έπιπλα” in Greek), and many others. To determine the website’s ranking for these keywords, searches were conducted in Google’s search engine. It’s worth noting that searches conducted within Greece may yield location-specific results. To address this, searches were performed using the Anonymouse.org site, was used for searches, which employs proxy servers to protect the location and personal data of searchers.

### **Structured Data**

Many users use images to make purchasing decisions. Thanks to the implementation of microdata application on the new website, every image on the website now appears in Google searches as a product, complete with its own price and title. This also applies to product listings, where the price is now visible, as depicted in **Figure 71.**

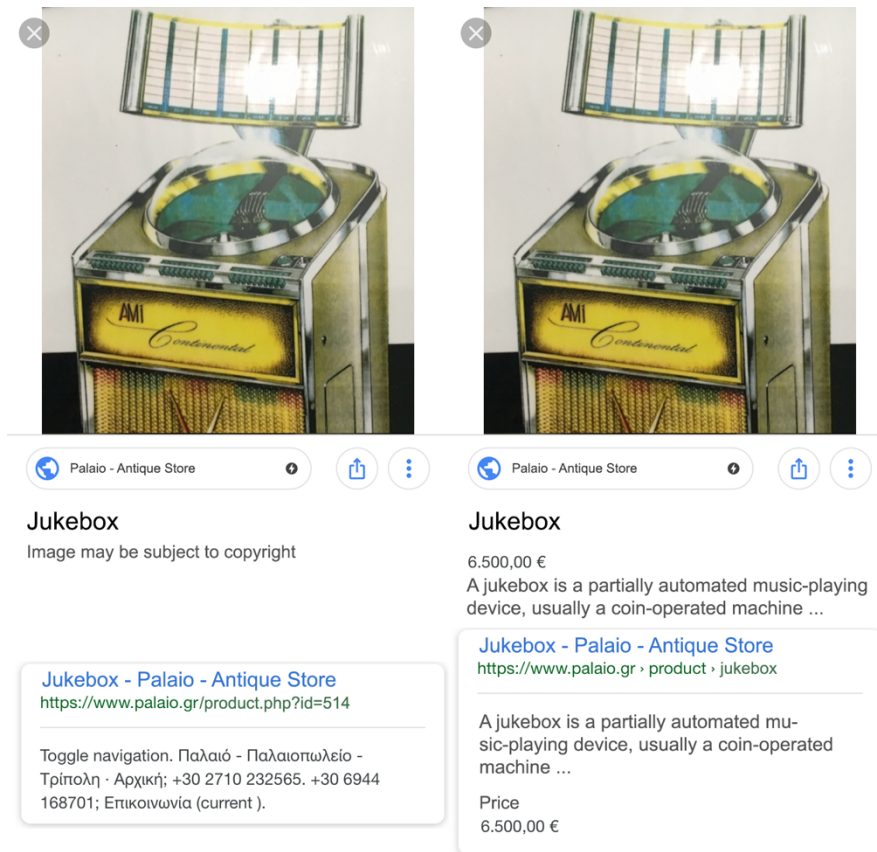


Figure 71. Structured Data Results.

### Accelerated Mobile Page

Following the implementation of AMP technologies, the website experienced an increase in visits from mobile device users. It was observed that the load time as well as the page size of the AMP page are significantly lower compared to those of the non-AMP website. According to Pingdom, the website speed reduced to 1.31s and the AMP page load time to 427ms.

### Open Graph Protocol

With the application of the Open Graph Protocol on the website [354], there was a substantial improvement in indexing.. **Figure 72** illustrates the difference in indexing between Meta Tags and OG Tags.

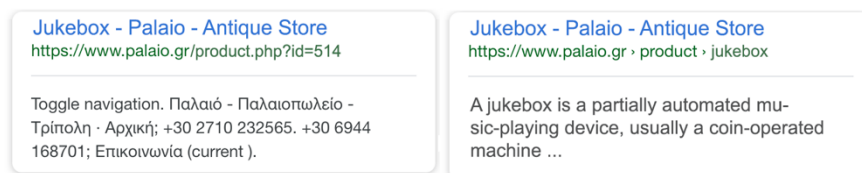
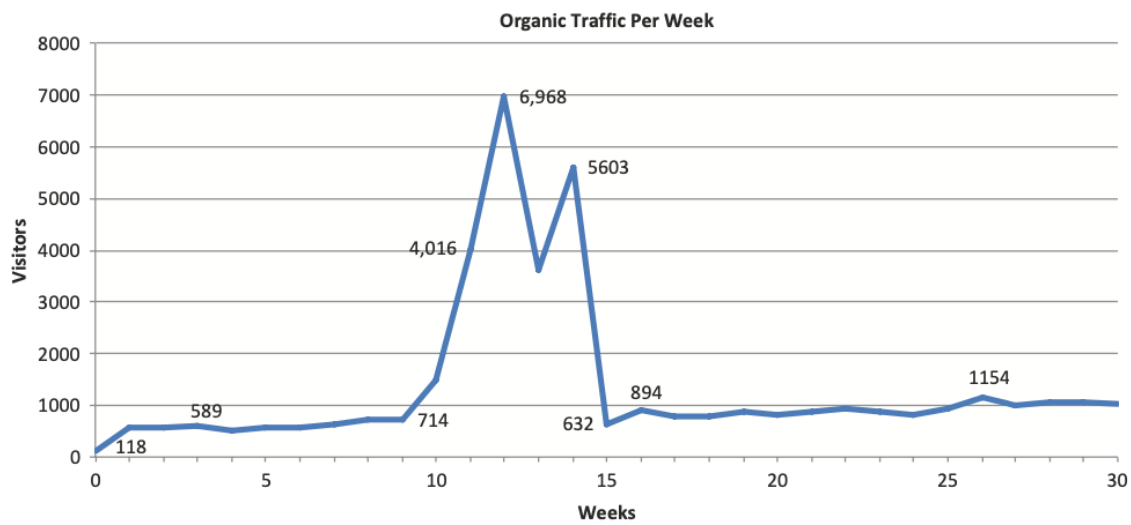


Figure 72. Indexing Results between Meta Tags and OG Tags

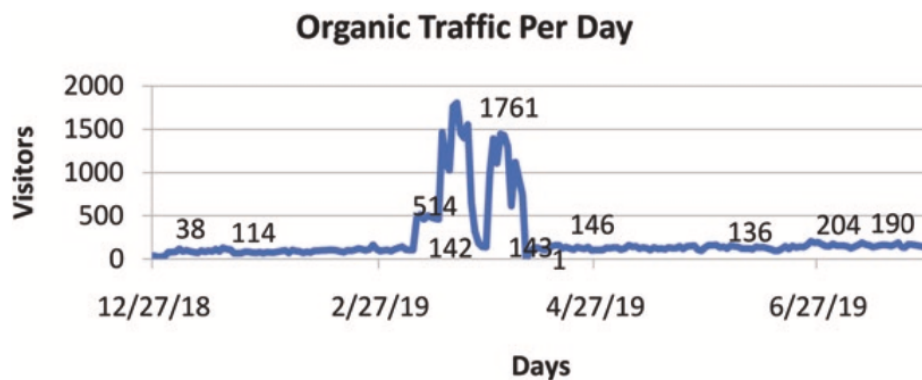
#### 4.11.4. Section Research Findings and Discussion

At the end of 2018, the website had 118 weekly visitors. Currently, it receives 1041 visits per week, indicating an impressive 782% increase in organic traffic following the implementation of SEO, as depicted in **Figure 73**. Weekly measurements provide a more representative perspective compared to daily measurements. Nevertheless, daily measurements are also presented below, to examine the variance in Organic Traffic in more detail.

Digging deeper, on December 27, 2018, the website had only 38 visitors. Seven months later, on July 27, 2019, the visitor count increased to 156 guests daily. This represents a remarkable 310% surge in organic traffic following the SEO implementation on the website as illustrated in **Figure 74**.



**Figure 73.** Organic traffic per week - statistics exported from google analytics



**Figure 74.** Organic traffic per day - statistics exported from google analytics

The substantial variation in traffic from week 9 to week 15 can be attributed to the implementation of Google Adwords advertising. At the conclusion of the 9th week, prior to the initiation of the Google Adwords promotion, the website had 714 weekly visitors. It was observed that, after the conclusion of the campaign at the end of the 15th week, the weekly visitor count dropped to 632. Contrary to the anticipated increase in organic traffic, following the paid advertising campaign, the website experienced a decrease to 632 visitors per week, reflecting an 11.48% decline in traffic. Subsequently, the organic traffic gradually increased once again.

#### 4.11.5. Section Conclusion

In a competitive environment, every website must enhance itself in order to attract more visitors. SEO plays a crucial role in helping these websites increase their organic traffic. To assess whether SEO techniques and technologies bring about real and effective results, they were applied to a live E-commerce platform. The results of a seven-month period were observed using the Google Analytics platform, and the final outcomes were extremely positive for all aspects, including organic traffic, page ranking, and user experience.

## 5. The Accelerated Mobile Pages (AMP) and its contribution to Web Traffic in the realm of E-commerce

*Adapted from:*

Roumeliotis, K.I.; Tselikas, N.D. *Accelerated Mobile Pages: A Comparative Study*. *International Conference on Business Intelligence and Modelling*. Springer Proceedings in Business and Economics. Springer, Cham 2021, 57–63. [https://doi.org/10.1007/978-3-030-57065-1\\_5](https://doi.org/10.1007/978-3-030-57065-1_5)

### 5.1. Section Summary and Significance of the Study

Over the last decade, the number of mobile subscriptions has rapidly grown, surpassing 7.7 billion by late 2017 [27]. Despite this rapid growth, the overall user experience for mobile users remains suboptimal. For instance, some web pages load slowly, and others do not utilize responsive design technologies. Following discussions with news publishers and technology companies, aimed at improving the performance of the mobile web, Google announced the AMP project in October 2015 [419].

AMP is an open-source HTML framework that provides a straightforward way to create web pages that are fast, smooth-loading, and prioritize user experience above everything else [347]. By May 2017, Google reported that 900,000 websites were publishing AMP pages, with more than two billion AMP pages published globally. Additionally, AMP usage reduces bandwidth consumption from image by 50% without compromising perceived quality [27].

AMP pages first appeared to web searchers in February 2016 when Google began showing AMP versions of webpages in SERPs. By September 2016, Google started linking to AMP content in the main mobile search results area. AMP pages are displayed in the same way as non-AMP pages. To distinguish between them, Google added a lightning strike icon before the search result [27].

In this chapter, we present how the AMP project works, the requirements to categorize a web page as an AMP, as well as the impact of AMP on webpage loading speed. Furthermore, we have implemented AMP technologies into a real-life E-commerce project and present a comparison between Organic and AMP traffic, respectively.

### 5.2. AMP Usage and Restrictions

#### 5.2.1. How AMP works

AMP, like any other web app, comprises of three components: HTML, CSS, and JavaScript code. However, AMP pages are subject to certain restrictions. Google bots constantly scan the web to discover and index new web pages to the search results - a process known as spidering. When a new web page is found, the bot crawls its content and uses it to rank the page in search results. In contrast, when a Google bot detects that a web page uses AMP technologies or has an AMP alternative page, it caches its source code on a Google server. If a Google user searches for a keyword on a mobile device, the Google search engine displays the AMP version of the website in the search results instead of the non-AMP page. Upon accessing the AMP page, the user receives the cached web page from the Google server, accompanied by a built-in validation system that ensures the web page's functionality [28]. **Figure 75** outlines the caching process.

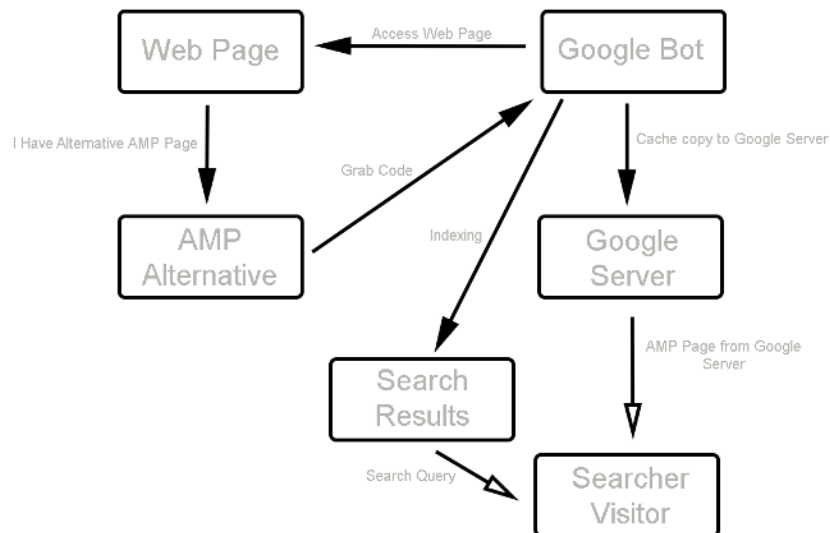


Figure 75. Procedure of caching.

### 5.2.2. AMP Restrictions and Components

To begin with, AMP pages consist of the AMP JS library, which is placed first after the HTML tag, along with built-in components, extended components, as well as custom scripts and styling components.

Regarding the built-in components, many HTML tags differ on an AMP page, resulting in some restrictions [25]. For instance, an image tag is defined as `<amp-img>` and the optional tags (`alt`, `width`, `height`, or `layout = "responsive"`) are required to display the image correctly.

The primary goal of the AMP project was to create a lightweight page without unnecessary code. Consequently, the AMP team created extended free components to introduce new styling and functionality features to the existing AMP code.

To load as quickly as possible, an AMP page does not have any external styling files embedded in its `<head>` HTML tag. All CSS scripts should be posted in-line within an `amp-custom <style>` tag, as illustrated in Figure 76 [26]. Conversely, there is the option to load external custom components—JavaScript files, such as the jQuery library with some restrictions (Figure 77). The JavaScript file is pre-cached by Google servers and can only be executed within the rendered area of the `amp-script` tag. This approach ensures that no custom JavaScript will slow down the web page, loading the `amp-script` after the overall loading of the AMP content [28].

```
<style amp-custom>body {background-color: white;} </style>
```

Figure 76. AMP in-line CSS styling.

```
<amp-script layout="container" src="https://domain.com/file.js">
  <p>The content can be modified by javascript.</p>
</amp-script>
```

Figure 77. AMP external JavaScript.

Every AMP page must have a different URL than the non-AMP page (Figure 78).

The non-AMP page needs to include a canonical link and an alternative AMP link tag in its header to be discovered by Google bot, as outlined in Figure 79. Conversely, the AMP page should contain the canonical link tag, directing back to the regular non-AMP version.

```

non-AMP Page: https://www.domain.com/article/seo
AMP Page: https://www.domain.com/article/seo/amp
or AMP Page: https://www.domain.com/amp/article/seo
or AMP Page: https://amp.domain.com/article/seo
or Google Server:
https://www.google.com/amp/s/www.domain.com/article/seo

```

Figure 78. AMP URL structure.

```

<link rel="canonical" href="https://www.domain.com/article/seo">
<link rel="amphtml" href="https://www.domain.com/article/seo/amp">

```

Figure 79. AMP html and canonical link tag.

### 5.3. AMP Impact to a Real-Life E-commerce

This section presents an original research involving the implementation of an AMP project on a real-life e-commerce platform. The statistical results for the 12-months period following the AMP implementation were collected from the Google Analytics platform.

#### 5.3.1. AMP Speed Performance

After creating the AMP page, we measured the load time for both AMP and non-AMP counterpart (Figure 80). The results demonstrated that the AMP load time is 82% less than that of the non-AMP, respectively. In practical terms, the website owner has saved more than 7.2 GB in traffic usage in a single year. This substantial difference is attributed to the built-in AMP JS library, ensuring the fast rendering of AMP HTML pages [25].

Non-AMP Webpage	Load time <b>1.67 s</b>	Page size <b>2.8 MB</b>
AMP Webpage	Load time <b>302 ms</b>	Page size <b>438.3 KB</b>

Figure 80. AMP vs non-AMP webpage speed.

#### 5.3.2. Comparing Organic Traffic with AMP Traffic

SEO is the science of optimizing web pages and content to increase organic traffic. SEO often involves making small modifications to various parts of the website [295]. While these individual modifications might not have a significant impact on organic traffic, combining SEO techniques with quality content certainly will [288].

On the other hand, AMP visitors contribute to AMP traffic, characterized by Google as referral traffic. According to Google's definition of referral, any external page that sends traffic to our website is counted as a referral.

A twelve-months comparative analysis between organic traffic and AMP traffic is presented in Figure 81.

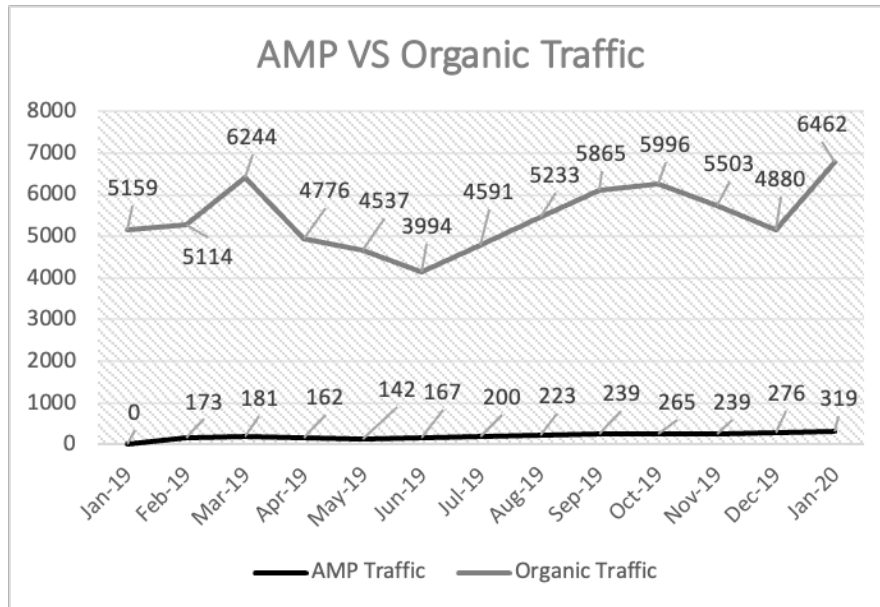


Figure 81. AMP vs organic traffic.

From the statistic results, we can conclude that AMP traffic is much lower than organic traffic. However, when combining the two, overall website traffic, and, for instance, e-commerce sales increased. The goal of every SEO campaign is to enhance website traffic while ensuring visitor satisfaction. The AMP project, with its mobile-friendliness and fast load time, can satisfy returning visitors and attract new organic ones.

An analysis of the results reveals that the AMP project has both pros and cons, as shown in Table 40. Even though the number of pros and cons is equal, the benefits of using the AMP project create valuable traffic, making the disadvantages less significant. On the other hand, the AMP platform should be more attractive to developers and provide them with additional features, such as the use of external style-sheets, to enhance the final appearance.

Table 40. Pros and Cons of using AMP project.

Pros	Cons
Less load time	It provides fewer results than Organic Traffic
Mobile-friendliness	Updating is required for two webpages when updates announcements are made
Better rankings on search results	Implementation incurs additional costs
More satisfied mobile users	Appearance and effects may be sacrificed for usability.
More Traffic and Potential Customers	AMP traffic is not considered actual organic traffic
Traffic and Web server resources saving	AMP traffic is the type that brings customers, not just visitors

#### 5.4. Section Conclusion

The e-commerce environment is evolving day by day, with new technologies continually emerging. To thrive in the fierce competition, e-commerce platforms must adapt and incorporate these innovations. The Google AMP project has been developed to assist e-commerce in attracting more mobile users and satisfying the returning ones. In this chapter we presented how AMP project works and outlined the specifications that a web page must adhere to, in order to be classified as AMP. We also implemented AMP techniques on a live e-commerce web site, presented the results over a twelve-months period gathered from the Google Analytics platform, and concluded by outlining the corresponding pros and cons of utilizing the AMP project.

## **Conclusion**

This thesis explores strategies and technologies aimed at fostering a highly optimized, accessible, and sustainable web and E-commerce landscape. The investigation involved creating software tools in both PHP and Python programming languages, leveraging advanced LLMs and NLP to automate previously challenging e-commerce processes. To validate and amplify the findings, the research employed data analysis tools, predictive modeling techniques, and ML algorithms. The integration of these diverse technologies results in an E-commerce ecosystem poised not only to withstand but also to thrive in the face of future intense competition.

In **Chapter 2**, the exploration revolves around the seamless integration of LLMs into the realm of E-commerce, emphasizing the utilization of their advanced AI and NLP capabilities to automate various e-commerce processes. The initial inquiry focuses on NLP and its proficiency in comprehending human language, specifically within the context of the GPT Architecture. The study delves into the groundbreaking utilization of transformers and self-attention mechanisms during the pre-training phase, which involves the training of GPT models with billion parameters and their subsequent adaptability for fine-tuning in domain-specific tasks.

This study specifically scrutinized notable models such as GPT-3.5, GPT-4, and LLaMA-2, exploring their integration into E-commerce and the potential enhancements they could bring, particularly in terms of automation, with the aim of elevating both the functional and customer-centric aspects of e-commerce. It's noteworthy that the exploration extended beyond LLMs, encompassing renowned NLP models such as BERT and RoBERTa. Additionally, the study delved into unsupervised and supervised learning algorithms, including k-means clustering, content-based filtering (CBF), hierarchical clustering, as well as logistic regression and neural network algorithms.

The conducted studies covered diverse topics such as sentiment analysis, recommender systems, sustainable purchasing decisions, and churn modeling. Specifically, the research proposed leveraging well-known LLMs such as GPT-3.5 and LLaMA-2 models for product review evaluation. Following fine-tuning, a comparative analysis demonstrated the superiority of LLMs in performing sentiment analysis tasks. Furthermore, the research suggested employing GPT-4 as an evaluation tool for unsupervised models, highlighting its efficacy in discerning contextual meanings in recommender systems. Additionally, the study proposed utilizing GPT models to reveal sustainable product characteristics, thereby promoting sustainable purchase decisions. Finally, the research recommended employing advanced supervised learning algorithms like logistic regression and neural networks to identify potential churn among e-commerce customers.

The results obtained from chapter 2 were highly encouraging, showcasing the precision of LLMs in executing diverse tasks across different e-commerce domains, both prior to and after fine-tuning.

In **Chapter 3** of the exploration into Web Accessibility, a comprehensive analysis of the WCAG was conducted. This involved a meticulous examination of each design principle, guideline, checkpoint, and success criterion. The research revealed limitations within WCAG, encompassing both the Accessibility Guidelines themselves and the efforts of web developers, as well as shortcomings in machine auditing tools.

As the investigation progressed, the focus shifted towards PWAs, an emerging web app technology. PWAs utilize a combination of cutting-edge technologies, including service workers, app shell architecture, web app manifest, and caching storage API, to provide advanced offline functionality. Through a parallel critical evaluation, the study explored the advantages and limitations associated with adopting this innovative technology.

Integrating the insights gained from the assessments of Accessibility and PWAs, a meticulous selection was made of a sample of PWA websites and e-commerce platforms. Employing popular accessibility evaluation and auditing tools, a comparative analysis was manually conducted against

their non-PWA counterparts. Following sample collection and a descriptive analysis utilizing Jupyter and Python, valuable insights into the accessibility of PWAs when compared to traditional websites were derived.

This study underscores the pivotal role of accessibility in the web landscape, shedding light on the identified limitations within WCAG and presenting PWA as a promising avenue for addressing and advancing web accessibility standards.

In **Chapter 4**, the SLR extensively explored the domain of SEO, emphasizing its crucial role in establishing a robust online presence for e-commerce entities and attaining higher rankings in SERPs. Beyond traditional SEO methods, the research championed the integration of existing web technologies to elevate SEO practices.

To operationalize the research objectives, prototype tools were developed using PHP and Python, imbued with crawler-like functionalities. These tools meticulously analyzed targeted web pages, extracting valuable insights into deployed SEO techniques. To simulate real-world scenarios, these tools were interconnected with APIs, facilitating the retrieval of additional data such as backlink count, DA, keywords, search ranking positions, responsive design, and website speed.

Moreover, the sophistication of the software was enhanced by incorporating ML and predictive algorithms, resulting in the creation of a pre-trained model. This model played a pivotal role in predicting the requisite number of backlinks and DA needed for a website to secure a coveted first-page ranking in search results, accounting for the intricacies of its competitive landscape.

Importantly, these cutting-edge tools have been released as open-source on GitHub, democratizing access for non-specialized SEO teams and business owners. They offer a pragmatic means to implement SEO strategies and assess results through the software's recommendations.

To assess the effectiveness of these tools and gain practical insights into the efficiency of various SEO techniques, case studies were conducted on specific domains within live e-commerce platforms. One instance involved developing a Python-based software to uncover specific SEO techniques that e-commerce should adopt to outperform competitors, providing tangible recommendations. These strategies were then implemented on a live e-commerce website, and their impact on traffic and search engine ranking was monitored. Additionally, a case study examined the SEO techniques utilized by airlines to enhance sustainability within their competitive landscape, highlighting practices often overlooked despite existing SEO guidelines. Similar methodologies were employed in analyzing SEO strategies on hotel websites, revealing valuable insights into their utilization and effectiveness.

Furthermore, a separate case study applied well-known SEO techniques to a live e-commerce platform, allowing for the capture and evaluation of results over a month. This comprehensive approach significantly contributes to enhancing the understanding of SEO practices and offers practical tools for optimizing online presence within the broader community.

In **Chapter 5**, attention turned to the burgeoning growth of the mobile phone industry, prompting webmasters to explore alternative methods for developing web apps that cater to mobile-friendly experiences. Notably, Google's AMP emerged as a technology capturing the interest of webmasters. By leveraging existing technologies and imposing specific restrictions, AMP aimed to generate pre-loaded web pages, offering advantages in terms of reduced load time. These pages not only boasted a decrease in page size but also promised improved rankings in search results, leading to increased organic traffic owing to their enhanced speed and user-friendliness.

In this study, the focus was on verifying whether AMP indeed delivered the promised speed and ranking benefits. To achieve this, a comparative analysis was conducted on the pages of a live e-commerce site, juxtaposing them against the corresponding pages of the same e-commerce site created using AMP technology. This analysis aimed to provide insights into the effectiveness of AMP in delivering the anticipated advantages for mobile-friendly web applications.

Throughout my extensive research, the primary goal was to discover and refine practical methods. This involved seamlessly integrating cutting-edge technologies. My overarching objective was to equip E-commerce not only to navigate but also to excel in the face of evolving and intensifying future competition. Staying at the forefront of technological advancements played a crucial role in

developing innovative and applied solutions that enhance the resilience and competitiveness of E-commerce platforms.

Undoubtedly, research is an ongoing and iterative process, highlighting dedication to adaptability and staying updated on emerging trends. As I navigate the dynamic landscape of E-commerce, I am optimistic about the potential to make significant contributions through both current and future initiatives. I look forward to the opportunity to delve deeper into my findings, refining strategies, and making meaningful contributions that will shape the future of E-commerce. My aspiration is to actively participate in the continuous evolution of this dynamic industry, ensuring its sustained growth and success.

## **Bibliography**

1. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in E-Commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation. *Natural Language Processing Journal* **2024**, *6*, 100056, doi:10.1016/J.NLP.2024.100056.
2. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* **2023**, *Vol. 15*, Page 192 **2023**, *15*, 192, doi:10.3390/FI15060192.
3. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Precision-Driven Product Recommendation Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems. *Software* **2024**, *Vol. 3*, Pages 62-80 **2024**, *3*, 62–80, doi:10.3390/SOFTWARE3010004.
4. Ecommerce Customer Churn Analysis and Prediction Available online: <https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction> (accessed on 12 February 2024).
5. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models. *Meta-Radiology* **2023**, *1*, 100017, doi:10.1016/J.METRAD.2023.100017.
6. Maroto-Gómez, M.; Castro-González, Á.; Castillo, J.C.; Malfaz, M.; Salichs, M.Á. An Adaptive Decision-Making System Supported on User Preference Predictions for Human–Robot Interactive Communication. *User Model User-adapt Interact* **2023**, *33*, 359–403, doi:10.1007/S11257-022-09321-2/TABLES/5.
7. Liu, Y.; Zhang, W.-N.; Chen, Y.; Zhang, Y.; Bai, H.; Feng, F.; Cui, H.; Li, Y.; Che, W. Conversational Recommender System and Large Language Model Are Made for Each Other in E-Commerce Pre-Sales Dialogue. **2023**.
8. Roumeliotis, K.I.; Tselikas, N.D. Evaluating Progressive Web App Accessibility for People with Disabilities. *Network* **2022**, *2*, doi:10.3390/network2020022.
9. Hamid, S.; Bawany, N.Z.; Zahoor, K. Assessing Ecommerce Websites: Usability and Accessibility Study. *2020 International Conference on Advanced Computer Science and Information Systems, ICACIS 2020* **2020**, 199–204, doi:10.1109/ICACIS51025.2020.9263162.
10. Broderick, A. Of Rights and Obligations: The Birth of Accessibility. *The International Journal of Human Rights* **2020**, *24*, 393–413, doi:10.1080/13642987.2019.1634556.
11. Sohaib, O.; Kang, K. E-Commerce Web Accessibility for People with Disabilities. *Lecture Notes in Information Systems and Organisation* **2017**, *22*, 87–100, doi:10.1007/978-3-319-52593-8\_6/COVER.
12. Sheppard, D. Introduction to Progressive Web Apps. *Beginning Progressive Web App Development* **2017**, 3–10, doi:10.1007/978-1-4842-3090-9\_1.
13. Mhaske, A.; Bhattad, A.; Khamkar, P.; More, R. Progressive Web App for Educational System. *International Research Journal of Engineering and Technology* **2018**.
14. (PDF) Impact of Progressive Web Apps on Web App Development Available online: [https://www.researchgate.net/publication/330834334\\_Impact\\_of\\_Progressive\\_Web\\_Apps\\_on\\_Web\\_App\\_Development](https://www.researchgate.net/publication/330834334_Impact_of_Progressive_Web_Apps_on_Web_App_Development) (accessed on 9 November 2023).
15. Progressive Web Apps | Web.Dev Available online: <https://web.dev/explore/progressive-web-apps> (accessed on 9 November 2023).

16. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Airlines' Sustainability Study Based on Search Engine Optimization Techniques and Technologies. *Sustainability (Switzerland)* **2022**, *14*, doi:10.3390/su141811225.
17. Roumeliotis, K.I.; Tselikas, N.D.; Tryfonopoulos, C. Greek Hotels' Web Traffic: A Comparative Study Based on Search Engine Optimization Techniques and Technologies. *Digital* **2022**, *2*, doi:10.3390/digital2030021.
18. Roumeliotis, K.I.; Tselikas, N.D. Search Engine Optimization Techniques: The Story of an Old-Fashioned Website. In Proceedings of the Springer Proceedings in Business and Economics; 2021.
19. Ologunbe, J.; Taiwo, E.O. The Importance of SEO and SEM in Improving Brand Visibility in E-Commerce Industry; A Study of Decathlon, Amazon and ASOS. *SSRN Electronic Journal* **2023**, doi:10.2139/SSRN.4638890.
20. Roumeliotis, K.I.; Tselikas, N.D. An Effective SEO Techniques and Technologies Guide-Map. *Journal of Web Engineering* **2022**, *21*, doi:10.13052/jwe1540-9589.21510.
21. Park, M. SEO for an Open Access Scholarly Information System to Improve User Experience. *Inf Discov Deliv* **2018**, *46*, 77–82, doi:10.1108/IDD-08-2017-0060/FULL/XML.
22. Roumeliotis, K.I.; Tselikas, N.D. Accelerated Mobile Pages: A Comparative Study. In Proceedings of the Springer Proceedings in Business and Economics; 2021.
23. AMP - a Web Component Framework to Easily Create User-First Web Experiences - Amp.Dev Available online: <https://amp.dev/> (accessed on 9 November 2023).
24. Make Your Pages Discoverable - Amp.Dev Available online: <https://amp.dev/documentation/guides-and-tutorials/optimize-and-measure/discovery> (accessed on 9 November 2023).
25. Revji, D.G.; Narsing, D.R.; Dendge, M. Accelerated Mobile Pages AMP. *International Journal of Trend in Scientific Research and Development* **2018**, *Volume-2*, 338–341, doi:10.31142/IJTSRD15821.
26. (PDF) Impact of Accelerated Mobile Pages Format on Corporate Web Sites Available online: [https://www.researchgate.net/publication/318722347\\_Impact\\_of\\_accelerated\\_mobile\\_pages\\_format\\_on\\_corporate\\_web\\_sites](https://www.researchgate.net/publication/318722347_Impact_of_accelerated_mobile_pages_format_on_corporate_web_sites) (accessed on 29 November 2023).
27. Jun, B.; Bustamante, F.E.; Whang, S.Y.; Bischof, Z.S. AMP up Your Mobile Web Experience: Characterizing the Impact of Google's Accelerated Mobile Project. *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM* **2019**, doi:10.1145/3300061.3300137.
28. Wibowo, A.; Aryotejo, G.; Mufadhol, M. Accelerated Mobile Pages from JavaScript as Accelerator Tool for Web Service on E-Commerce in the E-Business. *International Journal of Electrical and Computer Engineering (IJECE)* **2018**, *8*, 2399–2405, doi:10.11591/IJECE.V8I4.PP2399-2405.
29. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Unveiling Sustainability in Ecommerce: GPT-Powered Software for Identifying Sustainable Product Features. *Sustainability* **2023**, *Vol. 15, Page 12015* **2023**, *15*, 12015, doi:10.3390/SU151512015.
30. Rothman, D. *Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT*; 2021;
31. Hirschle Deep Natural Language Processing: Einstieg... - Google Scholar Available online: [https://scholar.google.com/scholar?hl=en&as\\_sdt=0%2C5&q=Hirschle+Deep+Natural+Language+Processing%3A+Einstieg+in+Word+Embedding&btnG=](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Hirschle+Deep+Natural+Language+Processing%3A+Einstieg+in+Word+Embedding&btnG=) (accessed on 14 November 2023).
32. Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv Neural Inf Process Syst* **2017**, *30*.
33. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* **2019**, 2978–2988, doi:10.18653/v1/p19-1285.

34. Liu, Z.; Yu, X.; Zhang, L.; Wu, Z.; Cao, C.; Dai, H.; Zhao, L.; Liu, W.; Shen, D.; Li, Q.; et al. DeID-GPT: Zero-Shot Medical Text De-Identification by GPT-4. **2023**.
35. Ray, P.P. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. *Internet of Things and Cyber-Physical Systems* **2023**, *3*, 121–154, doi:10.1016/J.IOTCPS.2023.04.003.
36. Overview of Game-Changing ChatGPT: AI Power to Be Feared or Embraced? Available online: <https://hospitalityinsights.ehl.edu/chatgpt-overview> (accessed on 14 November 2023).
37. Kheiri, K.; arXiv:2307.10234, H.K. preprint; 2023, undefined Sentimentgpt: Exploiting Gpt for Advanced Sentiment Analysis and Its Departure from Current Machine Learning. *arxiv.orgK Kheiri, H KarimiarXiv preprint arXiv:2307.10234, 2023•arxiv.org*.
38. Troya, A.; Pillai, R.G.; Rivero, C.R.; Genc, Z.; Kayal, S.; Araci, D. Aspect-Based Sentiment Analysis of Social Media Data with Pre-Trained Language Models. *ACM International Conference Proceeding Series* **2021**, 8–17, doi:10.1145/3508230.3508232.
39. Cayamcela, M.E.M.; Lim, W. Fine-Tuning a Pre-Trained Convolutional Neural Network Model to Translate American Sign Language in Real-Time. *2019 International Conference on Computing, Networking and Communications, ICNC 2019* **2019**, 100–104, doi:10.1109/ICCNC.2019.8685536.
40. Singh, S.; Mahmood, A. The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access* **2021**, *9*, 68675–68702, doi:10.1109/ACCESS.2021.3077350.
41. Zhang, M.; Li, J. A Commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research* **2021**, *1*, 831–833, doi:10.1016/J.FMRE.2021.11.011.
42. Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach (Dordr)* **2020**, *30*, 681–694, doi:10.1007/S11023-020-09548-1/FIGURES/5.
43. Si, C.; Gan, Z.; Yang, Z.; Wang, S.; Wang, J.; Boyd-Graber, J.; Wang, L. Prompting Gpt-3 to Be Reliable. *arxiv.orgC Si, Z Gan, Z Yang, S Wang, J Wang, J Boyd-Graber, L WangarXiv preprint arXiv:2210.09150, 2022•arxiv.org*.
44. Wang, S.; Jin, P. A Brief Summary of Prompting in Using GPT Models. **2023**, doi:10.32388/IMZIQ.
45. Møller, A.; Dalsgaard, J.; Pera, A.; arXiv, L.A. preprint; 2023, undefined Is a Prompt and a Few Samples All You Need? Using GPT-4 for Data Augmentation in Low-Resource Classification Tasks. *arxiv.orgAG Møller, JA Dalsgaard, A Pera, LM AielloarXiv preprint arXiv:2304.13861, 2023•arxiv.org*.
46. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **2023**.
47. LLaMA 2: A Model Overview and Demo Tutorial with Paperspace Gradient Available online: <https://blog.paperspace.com/llama-2/> (accessed on 15 November 2023).
48. Meta-Llama/Llama-2-7b · Hugging Face Available online: <https://huggingface.co/meta-llama/Llama-2-7b> (accessed on 15 November 2023).
49. GitHub - Facebookresearch/Llama: Inference Code for LLaMA Models Available online: <https://github.com/facebookresearch/llama/tree/main> (accessed on 15 November 2023).
50. Models - Hugging Face Available online: <https://huggingface.co/models?other=llama-2> (accessed on 15 November 2023).
51. Llama 2 - Resource Overview - Meta AI Available online: <https://ai.meta.com/resources/models-and-libraries/llama/> (accessed on 15 November 2023).
52. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers* **2015**, *3*, 1715–1725, doi:10.18653/v1/p16-1162.
53. arXiv:2002.05202, N.S. preprint; 2020, undefined Glu Variants Improve Transformer. *arxiv.orgN ShazeerarXiv preprint arXiv:2002.05202, 2020•arxiv.org* **2020**.
54. Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; Wang, H. Preference Ranking Optimization for Human Alignment. *arxiv.orgF Song, B Yu, M Li, H Yu, F Huang, Y Li, H WangarXiv preprint arXiv:2306.17492, 2023•arxiv.org*.

55. Wölfel, M.; Taecharungroj, V. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing* 2023, Vol. 7, Page 35 **2023**, 7, 35, doi:10.3390/BDCC7010035.
56. Franckowiak, A.; Santander, M.; Sotnikov, V.; Chaikova, A. Language Models for Multimessenger Astronomy. *Galaxies* 2023, Vol. 11, Page 63 **2023**, 11, 63, doi:10.3390/GALAXIES11030063.
57. ECommerce - Worldwide | Statista Market Forecast Available online: <https://www.statista.com/outlook/dmo/ecommerce/worldwide#revenue> (accessed on 9 October 2023).
58. Li, L.; Yuan, L.; Tian, J. Influence of Online E-Commerce Interaction on Consumer Satisfaction Based on Big Data Algorithm. *Heliyon* **2023**, 9, doi:10.1016/j.heliyon.2023.e18322.
59. Wang, Y.; Lu, X.; Tan, Y. Impact of Product Attributes on Customer Satisfaction: An Analysis of Online Reviews for Washing Machines. *Electron Commer Res Appl* **2018**, 29, 1–11, doi:10.1016/J.ELERAP.2018.03.003.
60. Engler, T.H.; Winter, P.; Schulz, M. Understanding Online Product Ratings: A Customer Satisfaction Model. *Journal of Retailing and Consumer Services* **2015**, 27, 113–120, doi:10.1016/J.JRETCONSER.2015.07.010.
61. Liu, Y.; Wan, Y.; Shen, X.; Ye, Z.; Wen, J. Product Customer Satisfaction Measurement Based on Multiple Online Consumer Review Features. *Information* 2021, Vol. 12, Page 234 **2021**, 12, 234, doi:10.3390/INFO12060234.
62. Wang, Y.; Li, X. Mining Product Reviews for Needs-Based Product Configurator Design: A Transfer Learning-Based Approach. *IEEE Trans Industr Inform* **2021**, 17, 6192–6199, doi:10.1109/TII.2020.3043315.
63. Kang, H.; Yoo, S.J.; Han, D. Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews. *Expert Syst Appl* **2012**, 39, 6000–6010, doi:10.1016/J.ESWA.2011.11.107.
64. Li, Q.; Zhao, S.; Zhao, S.; Wen, J. Logistic Regression Matching Pursuit Algorithm for Text Classification. *Knowl Based Syst* **2023**, 277, 110761, doi:10.1016/J.KNOSYS.2023.110761.
65. Ahmad, S.N.; Laroche, M. Extracting Marketing Information from Product Reviews: A Comparative Study of Latent Semantic Analysis and Probabilistic Latent Semantic Analysis. *Journal of Marketing Analytics* **2023**, 1–15, doi:10.1057/S41270-023-00218-6/METRICS.
66. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv Neural Inf Process Syst* **2020**, 2020-December.
67. Kumar, V.; Ayodeji, O.G. E-Retail Factors for Customer Activation and Retention: An Empirical Study from Indian e-Commerce Customers. *Journal of Retailing and Consumer Services* **2021**, 59, 102399, doi:10.1016/J.JRETCONSER.2020.102399.
68. Murali, S.; Pugazhendhi, S.; Muralidharan, C. Modelling and Investigating the Relationship of after Sales Service Quality with Customer Satisfaction, Retention and Loyalty – A Case Study of Home Appliances Business. *Journal of Retailing and Consumer Services* **2016**, 30, 67–83, doi:10.1016/J.JRETCONSER.2016.01.001.
69. Kassim, N.; Abdullah, N.A. Customer Loyalty in E-Commerce Settings: An Empirical Study. *Electronic Markets* **2008**, 18, 275–290, doi:10.1080/10196780802265843.
70. Meire, M. Customer Comeback: Empirical Insights into the Drivers and Value of Returning Customers. *J Bus Res* **2021**, 127, 193–205, doi:10.1016/J.JBUSRES.2021.01.017.
71. Rosli, N.; Nayan, S.M. Why Customer First? *Journal of Undergraduate Social Science and Technology* **2020**, 2, doi:10.1108/JABS-01-2019-0020.
72. Utz, S.; Kerkhof, P.; Van Den Bos, J. Consumers Rule: How Consumer Reviews Influence Perceived Trustworthiness of Online Stores. *Electron Commer Res Appl* **2012**, 11, 49–58, doi:10.1016/J.ELERAP.2011.07.010.

73. Roethke, K.; Klumpe, J.; Adam, M.; Benlian, A. Social Influence Tactics in E-Commerce Onboarding: The Role of Social Proof and Reciprocity in Affecting User Registrations. *Decis Support Syst* **2020**, *131*, 113268, doi:10.1016/J.DSS.2020.113268.
74. Pei, X.L.; Guo, J.N.; Wu, T.J.; Zhou, W.X.; Yeh, S.P. Does the Effect of Customer Experience on Customer Satisfaction Create a Sustainable Competitive Advantage? A Comparative Study of Different Shopping Situations. *Sustainability* **2020**, *Vol. 12*, Page 7436 **2020**, *12*, 7436, doi:10.3390/SU12187436.
75. Griva, A. "I Can Get No e-Satisfaction". What Analytics Say? Evidence Using Satisfaction Data from e-Commerce. *Journal of Retailing and Consumer Services* **2022**, *66*, 102954, doi:10.1016/J.JRETCONSER.2022.102954.
76. Sundararaj, V.; Rejeesh, M.R. A Detailed Behavioral Analysis on Consumer and Customer Changing Behavior with Respect to Social Networking Sites. *Journal of Retailing and Consumer Services* **2021**, *58*, 102190, doi:10.1016/J.JRETCONSER.2020.102190.
77. Dwidienawati, D.; Tjahjana, D.; Abdinagoro, S.B.; Gandasari, D.; Munawaroh Customer Review or Influencer Endorsement: Which One Influences Purchase Intention More? *Heliyon* **2020**, *6*, doi:10.1016/j.heliyon.2020.e05543.
78. Askalidis, G.; Kim, S.J.; Malthouse, E.C. Understanding and Overcoming Biases in Online Review Systems. *Decis Support Syst* **2017**, *97*, 23–30, doi:10.1016/j.dss.2017.03.002.
79. Guha Majumder, M.; Dutta Gupta, S.; Paul, J. Perceived Usefulness of Online Customer Reviews: A Review Mining Approach Using Machine Learning & Exploratory Data Analysis. *J Bus Res* **2022**, *150*, 147–164, doi:10.1016/J.JBUSRES.2022.06.012.
80. Ramaswamy, S.; DeClerck, N. Customer Perception Analysis Using Deep Learning and NLP. *Procedia Comput Sci* **2018**, *140*, 170–178, doi:10.1016/J.PROCS.2018.10.326.
81. Stephen, A.T. The Role of Digital and Social Media Marketing in Consumer Behavior. *Curr Opin Psychol* **2016**, *10*, 17–21, doi:10.1016/J.COPSYC.2015.10.016.
82. Liu, Q. (Ben); Karahanna, E.; Watson, R.T. Unveiling User-Generated Content: Designing Websites to Best Present Customer Reviews. *Bus Horiz* **2011**, *54*, 231–240, doi:10.1016/J.BUSHOR.2011.01.004.
83. Kanaan, A.G.; Wahsheh, F.R.; El-Ebiary, Y.A.B.; Wan Hamzah, W.M.A.F.; Pandey, B.; P, S.N. An Evaluation and Annotation Methodology for Product Category Matching in E-Commerce Using GPT. *2023 International Conference on Computer Science and Emerging Technologies (CSET)* **2023**, 1–6, doi:10.1109/CSET58993.2023.10346684.
84. Chen, J.; Tech, W.G.; Cho, J.H.D.; Ma, L.; Li, X.; Thakurdesai, N.; Xu, J.; Nag, K.; Korpeoglu, E.; Kumar, S. Knowledge Graph Completion Models Are Few-Shot Learners: An Empirical Study of Relation Labeling in E-Commerce with LLMs. **2023**.
85. Gao, D.; Chen, K.; Chen, B.; Dai, H.; Jin, L.; Jiang, W.; Ning, W.; Yu, S.; Xuan, Q.; Cai, X.; et al. LLMs-Based Machine Translation for E-Commerce., doi:10.2139/SSRN.4682559.
86. Wang, H.; Na, T. Rethinking E-Commerce Search. **2023**.
87. Al Wahshat, H.; Abu-ulbeh, W.; Yusoff, M.H.; Zakaria, M.D.; Amir Fazamin Wan Hamzah, W.M.; P, S.N. The Detection of E-Commerce Manipulated Reviews Using GPT-4. *2023 International Conference on Computer Science and Emerging Technologies (CSET)* **2023**, 1–6, doi:10.1109/CSET58993.2023.10346848.
88. Zhou, J.; Liu, B.; Hong, J.N.A.Y.; Lee, K.; Wen, M. Leveraging Large Language Models for Enhanced Product Descriptions in ECommerce. **2023**.
89. Shi, K.; Sun, X.; Wang, D.; Fu, Y.; Xu, G.; Li, Q. LLaMA-E: Empowering E-Commerce Authoring with Multi-Aspect Instruction Following. *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)* **2023**, *1*, doi:XXXXXXXX.XXXXXXX.
90. Ma, S.; Huang, S.; Huang, S.; Wang, X.; Li, Y.; Zheng, H.-T.; Xie, P.; Huang, F.; Jiang, Y. EcomGPT-CT: Continual Pre-Training of E-Commerce Large Language Models with Semi-Structured Data. **2023**.

91. Maragheh, R.Y.; Morishetti, L.; Giahi, R.; Nag, K.; Xu, J.; Cho, J.; Korpeoglu, E.; Kumar, S.; Achan, K. LLM-Based Aspect Augmentations for Recommendation Systems. **2023**.
92. Orzoł, M.; Szopik-Depczyńska, K. ChatGPT as an Innovative Tool for Increasing Sales in Online Stores. *Procedia Comput Sci* **2023**, *225*, 3450–3459, doi:10.1016/J.PROCS.2023.10.340.
93. Li, Y.; Ma, S.; Wang, X.; Huang, S.; Jiang, C.; Zheng, H.-T.; Xie, P.; Huang, F.; Jiang, Y. EcomGPT: Instruction-Tuning Large Language Models with Chain-of-Task Tasks for E-Commerce. **2023**.
94. GitHub - Kroumeliotis/Fine-Tuning-Gpt3.5-Llama2-for-Product-Reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews Available online: <https://github.com/kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews> (accessed on 16 November 2023).
95. Sklearn.Model\_selection.Train\_test\_split – Scikit-Learn 1.3.2 Documentation Available online: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.train\\_test\\_split.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) (accessed on 17 December 2023).
96. Zhang, K.; Zhou, F.; Wu, L.; Xie, N.; He, Z. Semantic Understanding and Prompt Engineering for Large-Scale Traffic Data Imputation. *Information Fusion* **2024**, *102*, 102038, doi:10.1016/J.INFFUS.2023.102038.
97. What Runs ChatGPT? Inside Microsoft’s AI Supercomputer | Featuring Mark Russinovich - YouTube Available online: <https://www.youtube.com/watch?v=Rk3nTUfRZmo> (accessed on 17 December 2023).
98. Kroumeliotis/Ecommerce-Reviews5029 – Run with an API on Replicate Available online: <https://replicate.com/kroumeliotis/ecommerce-reviews5029> (accessed on 17 December 2023).
99. Kroumeliotis/Ecommerce-Reviews50 – Run with an API on Replicate Available online: <https://replicate.com/kroumeliotis/ecommerce-reviews50> (accessed on 17 December 2023).
100. Bert-Base-Uncased · Hugging Face Available online: <https://huggingface.co/bert-base-uncased> (accessed on 17 December 2023).
101. BERT – Transformers 3.0.2 Documentation Available online: [https://huggingface.co/transformers/v3.0.2/model\\_doc/bert.html#berttokenizer](https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#berttokenizer) (accessed on 17 December 2023).
102. Pretrained Models – Transformers 3.3.0 Documentation Available online: [https://huggingface.co/transformers/v3.3.1/pretrained\\_models.html](https://huggingface.co/transformers/v3.3.1/pretrained_models.html) (accessed on 17 December 2023).
103. BERT – Transformers 3.0.2 Documentation Available online: [https://huggingface.co/transformers/v3.0.2/model\\_doc/bert.html#bertforsequenceclassification](https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#bertforsequenceclassification) (accessed on 17 December 2023).
104. Tqdm · PyPI Available online: <https://pypi.org/project/tqdm/> (accessed on 17 December 2023).
105. Roberta-Base · Hugging Face Available online: <https://huggingface.co/roberta-base> (accessed on 17 December 2023).
106. RoBERTa – Transformers 2.9.1 Documentation Available online: [https://huggingface.co/transformers/v2.9.1/model\\_doc/roberta.html#robertaforsequenceclassification](https://huggingface.co/transformers/v2.9.1/model_doc/roberta.html#robertaforsequenceclassification) (accessed on 17 December 2023).
107. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst* **2023**, *47*, 1–5, doi:10.1007/S10916-023-01925-4/TABLES/2.
108. Ai, Q.; Bai, T.; Cao, Z.; Chang, Y.; Chen, J.; Chen, Z.; Cheng, Z.; Dong, S.; Dou, Z.; Feng, F.; et al. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* **2023**, *4*, 80–90, doi:10.1016/J.AIOPEN.2023.08.001.
109. Cai, X.; Xiao, M.; Ning, Z.; Zhou, Y. Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-Based Data Augmentation. **2023**.
110. Yao, B.; Chen, G.; Zou, R.; Lu, Y.; Li, J.; Zhang, S.; Liu, S.; Hendler, J.; Wang, D. More Samples or More Prompt Inputs? Exploring Effective In-Context Sampling for LLM Few-Shot Prompt Engineering. **2023**.

111. Agarwal, A.; Chakraborty, M.; Chowdary, C.R. Does Order Matter? Effect of Order in Group Recommendation. *Expert Syst Appl* **2017**, *82*, 115–127, doi:10.1016/J.ESWA.2017.03.069.
112. Kalyan, K.S. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *Natural Language Processing Journal* **2023**, 100048, doi:10.1016/j.nlp.2023.100048.
113. Products Suggestions for WooCommerce – WordPress Plugin | WordPress.Org Available online: <https://wordpress.org/plugins/cart-products-suggestions-for-woocommerce/> (accessed on 18 February 2024).
114. Lai, C.H.; Tseng, K.C. Applying Deep Learning Models to Analyze Users’ Aspects, Sentiment, and Semantic Features for Product Recommendation. *Applied Sciences* **2022**, *Vol. 12*, Page 2118 **2022**, *12*, 2118, doi:10.3390/APP12042118.
115. Hell, F.; Taha, Y.; Hinz, G.; Heibei, S.; Müller, H.; Knoll, A. Graph Convolutional Neural Network for a Pharmacy Cross-Selling Recommender System. *Information* **2020**, *Vol. 11*, Page 525 **2020**, *11*, 525, doi:10.3390/INFO11110525.
116. Ghoshal, A.; Mookerjee, V.S.; Sarkar, S. Recommendations and Cross-Selling: Pricing Strategies When Personalizing Firms Cross-Sell. *Journal of Management Information Systems* **2021**, *38*, 430–456, doi:10.1080/07421222.2021.1912930.
117. Vatavwala, S.; Kumar, B.; Sharma, A. Enhancing Upselling and Cross-Selling in Business-to-Business Markets: The Critical Need to Integrate Customer Service and Sales Functions. *Customer Centric Support Services in the Digital Age* **2024**, 199–216, doi:10.1007/978-3-031-37097-7\_9.
118. Lesage, L.; Deaconu, M.; Lejay, A.; Meira, J.A.; Nichil, G.; State, R. A Recommendation System for Car Insurance. *Eur Actuar J* **2020**, *10*, 377–398, doi:10.1007/S13385-020-00236-Z/METRICS.
119. Park, C.H.; Yoon, T.J. The Dark Side of Up-Selling Promotions: Evidence from an Analysis of Cross-Brand Purchase Behavior☆. *Journal of Retailing* **2022**, *98*, 647–666, doi:10.1016/J.JRETAI.2022.03.005.
120. Zhu, T.; Harrington, P.; Li, J.; Tang, L. Bundle Recommendation in ECommerce. *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval* **2014**, 657–666, doi:10.1145/2600428.2609603.
121. Zhou, H.; Xiong, F.; Chen, H.A.; Zhou, H.; Xiong, F.; Chen, H. A Comprehensive Survey of Recommender Systems Based on Deep Learning. *Applied Sciences* **2023**, *Vol. 13*, Page 11378 **2023**, *13*, 11378, doi:10.3390/APP132011378.
122. Alcaraz-Herrera, H.; Cartlidge, J.; Toumpakari, Z.; Western, M.; Palomares, I. EvoRecSys: Evolutionary Framework for Health and Well-Being Recommender Systems. *User Model User-adapt Interact* **2022**, *32*, 883–921, doi:10.1007/S11257-021-09318-3/FIGURES/12.
123. Chen, J.; Chen, W.; Huang, J.; Fang, J.; Li, Z.; Liu, A.; Zhao, L. Co-Purchaser Recommendation for Online Group Buying. *Data Sci Eng* **2020**, *5*, 280–292, doi:10.1007/S41019-020-00138-W/TABLES/4.
124. Stöckli, D.R.; Khobzi, H. Recommendation Systems and Convergence of Online Reviews: The Type of Product Network Matters! *Decis Support Syst* **2021**, *142*, 113475, doi:10.1016/J.DSS.2020.113475.
125. Wijaya, I.W.R.; Mudjahidin Development of Conceptual Model to Increase Customer Interest Using Recommendation System in E-Commerce. *Procedia Comput Sci* **2022**, *197*, 727–733, doi:10.1016/J.PROCS.2021.12.194.
126. Aldino, A.A.; Pratiwi, E.D.; Setiawansyah; Sintaro, S.; Putra, A.D. Comparison of Market Basket Analysis to Determine Consumer Purchasing Patterns Using Fp-Growth and Apriori Algorithm. *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2021* **2021**, 29–34, doi:10.1109/ICOMITEE53461.2021.9650317.
127. Tewari, A.S. Generating Items Recommendations by Fusing Content and User-Item Based Collaborative Filtering. *Procedia Comput Sci* **2020**, *167*, 1934–1940, doi:10.1016/J.PROCS.2020.03.215.

128. Rahmatillah, I.; Astuty, E.; Sudirman, I.D. An Improved Decision Tree Model for Forecasting Consumer Decision in a Medium Groceries Store. *2023 IEEE 17th International Conference on Industrial and Information Systems, ICIIIS 2023 - Proceedings* **2023**, 245–250, doi:10.1109/ICIIIS58898.2023.10253592.
129. Jbene, M.; Tigani, S.; Rachid, S.; Chehri, A. Deep Neural Network and Boosting Based Hybrid Quality Ranking for E-Commerce Product Search. *Big Data and Cognitive Computing 2021, Vol. 5, Page 35* **2021**, 5, 35, doi:10.3390/BDCC5030035.
130. Telikani, A.; Gandomi, A.H.; Shahbahrami, A. A Survey of Evolutionary Computation for Association Rule Mining. *Inf Sci (N Y)* **2020**, 524, 318–352, doi:10.1016/J.INS.2020.02.073.
131. Wang, K.; Zhang, T.; Xue, T.; Lu, Y.; Na, S.G. E-Commerce Personalized Recommendation Analysis by Deeply-Learned Clustering. *J Vis Commun Image Represent* **2020**, 71, 102735, doi:10.1016/J.JVCIR.2019.102735.
132. Javed, U.; Javed, U.; Shaukat, K.; Hameed, I.A.; Iqbal, F.; Alam, T.M.; Luo, S. A Review of Content-Based and Context-Based Recommendation Systems. *International Journal of Emerging Technologies in Learning (ijET)* **2021**, 16, 274–306.
133. Isinkaye, F.O. Matrix Factorization in Recommender Systems: Algorithms, Applications, and Peculiar Challenges. *IETE J Res* **2023**, 69, 6087–6100, doi:10.1080/03772063.2021.1997357.
134. Wu, G.; Sanner, S.; Luo, K.; Soh, H. Deep Language-Based Critiquing for Recommender Systems. *RecSys 2019 - 13th ACM Conference on Recommender Systems* **2019**, 137–145, doi:10.1145/3298689.3347009.
135. Bhaskaran, S.; Marappan, R.; Santhi, B. Design and Analysis of a Cluster-Based Intelligent Hybrid Recommendation System for E-Learning Applications. *Mathematics 2021, Vol. 9, Page 197* **2021**, 9, 197, doi:10.3390/MATH9020197.
136. Ko, H.; Lee, S.; Park, Y.; Choi, A. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics* **2022**, 11, 141, doi:10.3390/ELECTRONICS11010141.
137. Bandyopadhyay, S.; Thakur, S.S.; Mandal, J.K. Product Recommendation for E-Commerce Business by Applying Principal Component Analysis (PCA) and K-Means Clustering: Benefit for the Society. *Innov Syst Softw Eng* **2021**, 17, 45–52, doi:10.1007/S11334-020-00372-5/METRICS.
138. Sinaga, K.P.; Yang, M.S. Unsupervised K-Means Clustering Algorithm. *IEEE Access* **2020**, 8, 80716–80727, doi:10.1109/ACCESS.2020.2988796.
139. Nguyen, L.V.; Nguyen, T.H.; Jung, J.J. Content-Based Collaborative Filtering Using Word Embedding: A Case Study on Movie Recommendation. *ACM International Conference Proceeding Series* **2020**, 96–100, doi:10.1145/3400286.3418253.
140. Sangaiah, A.K.; Javadpour, A.; Ja'fari, F.; Zhang, W.; Khaniabadi, S.M. Hierarchical Clustering Based on Dendrogram in Sustainable Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems* **2023**, 24, 15724–15739, doi:10.1109/TITS.2022.3222789.
141. Lamb, D.S.; Downs, J.; Reader, S. Space-Time Hierarchical Clustering for Identifying Clusters in Spatiotemporal Point Data. *ISPRS International Journal of Geo-Information* **2020**, 9, 85, doi:10.3390/IJGI9020085.
142. Li, T.; Rezaeipanah, A.; Tag El Din, E.S.M. An Ensemble Agglomerative Hierarchical Clustering Algorithm Based on Clusters Clustering Technique and the Novel Similarity Measurement. *Journal of King Saud University - Computer and Information Sciences* **2022**, 34, 3828–3842, doi:10.1016/J.JKSUCI.2022.04.010.
143. Amazon UK Products Dataset 2023 (2.2M Products) Available online: <https://www.kaggle.com/datasets/asaniczka/amazon-uk-products-dataset-2023> (accessed on 3 February 2024).
144. Sklearn.Model\_selection.Train\_test\_split — Scikit-Learn 1.4.0 Documentation Available online: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.train\\_test\\_split.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) (accessed on 3 February 2024).

145. GitHub - Kroumeliotis/Product-Recommendations-Software-Unsupervised-Models-Evaluated-by-GPT-4-LLM: Precision-Driven Product Recommendations Software: Un-Leashing the Power of Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems Available online: <https://github.com/kroumeliotis/Product-Recommendations-Software-Unsupervised-Models-Evaluated-by-GPT-4-LLM> (accessed on 3 February 2024).
146. Colab.Google Available online: <https://colab.google/> (accessed on 3 February 2024).
147. Walkowiak, T. Subject Classification of Texts in Polish - from TF-IDF to Transformers. **2021**, 457–465, doi:10.1007/978-3-030-76773-0\_44.
148. Kumar, V.; Subba, B. A Tfidfvectorizer and SVM Based Sentiment Analysis Framework for Text Data Corpus. *26th National Conference on Communications, NCC 2020* **2020**, doi:10.1109/NCC48643.2020.9056085.
149. Hong, Y.; Tantithamthavorn, C.; Thongtanunam, P.; Aleti, A. CommentFinder: A Simpler, Faster, More Accurate Code Review Comments Recommendation. *ESEC/FSE 2022 - Proceedings of the 30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering* **2022**, 507–519, doi:10.1145/3540250.3549119.
150. Hasani, A.M.; Singh, S.; Zahergivar, A.; Ryan, B.; Nethala, D.; Bravomontenegro, G.; Mendhiratta, N.; Ball, M.; Farhadi, F.; Malayeri, A. Evaluating the Performance of Generative Pre-Trained Transformer-4 (GPT-4) in Standardizing Radiology Reports. *Eur Radiol* **2023**, 1–9, doi:10.1007/S00330-023-10384-X/METRICS.
151. OpenAI API Available online: <https://openai.com/blog/openai-api> (accessed on 14 November 2023).
152. Litvaj, I.; Ponisciakova, O.; Stancekova, D.; Svobodova, J.; Mrazik, J. Decision-Making Procedures and Their Relation to Knowledge Management and Quality Management. *Sustainability* **2022**, Vol. 14, Page 572 **2022**, 14, 572, doi:10.3390/SU14010572.
153. White, M.A. Sustainability: I Know It When I See It. *Ecological Economics* **2013**, 86, 213–217, doi:10.1016/J.ECOLECON.2012.12.020.
154. Rondinelli, D.A.; Berry, M.A. Environmental Citizenship in Multinational Corporations: Social Responsibility and Sustainable Development. *European Management Journal* **2000**, 18, 70–84, doi:10.1016/S0263-2373(99)00070-5.
155. Purvis, B.; Mao, Y.; Robinson, D. Three Pillars of Sustainability: In Search of Conceptual Origins. *Sustain Sci* **2019**, 14, 681–695, doi:10.1007/S11625-018-0627-5/FIGURES/1.
156. Clark, G.; Kosoris, J.; Hong, L.N.; Crul, M. Design for Sustainability: Current Trends in Sustainable Product Design and Development. *Sustainability* **2009**, Vol. 1, Pages 409–424 **2009**, 1, 409–424, doi:10.3390/SU1030409.
157. Govindan, K. How Artificial Intelligence Drives Sustainable Frugal Innovation: A Multitheoretical Perspective. *IEEE Trans Eng Manag* **2022**, doi:10.1109/TEM.2021.3116187.
158. Peráček, T. E-Commerce and Its Limits in the Context of the Consumer Protection: The Case of the Slovak Republic., doi:10.24818/TBJ/2022/12/1.03.
159. Klimek, L.; Funta, R. Data and E-Commerce: An Economic Relationship. *Danube* **2021**, 12, 33–44, doi:10.2478/DANB-2021-0003.
160. Gupta, N.; Kumar, A. Deepika Store: A Dilemma of Being Sustainable or Profitable. *Socially Responsible Consumption and Marketing in Practice: Collection of Case Studies* **2022**, 113–129, doi:10.1007/978-981-16-6433-5\_9/COVER.
161. Gashi, F.; Peci, B. Protection of Personal Data and Privacy in Banking Sector in Kosovo and Its Impact in Consumer Protection. *Perspectives of Law and Public Administration* **2020**, 9.
162. Purcărea, T.; Ioan-Franc, V.; Ionescu, Ş.A.; Purcărea, I.M.; Purcărea, V.L.; Purcărea, I.; Mateescu-Soare, M.C.; Platon, O.E.; Orzan, A.O. Major Shifts in Sustainable Consumer Behavior in Romania and Retailers' Priorities in Agilely Adapting to It. *Sustainability* **2022**, Vol. 14, Page 1627 **2022**, 14, 1627, doi:10.3390/SU14031627.

163. Redondo, R.; Valor, C.; Carrero, I. Unraveling the Relationship between Well-Being, Sustainable Consumption and Nature Relatedness: A Study of University Students. *Appl Res Qual Life* **2022**, *17*, 913–930, doi:10.1007/S11482-021-09931-9/METRICS.
164. He, Y.; Li, X.; Huang, P.; Wang, J. Exploring the Road toward Environmental Sustainability: Natural Resources, Renewable Energy Consumption, Economic Growth, and Greenhouse Gas Emissions. *Sustainability* **2022**, *Vol. 14*, Page 1579 **2022**, *14*, 1579, doi:10.3390/SU14031579.
165. Taghvaei, V.M.; Nodehi, M.; Arani, A.A.; Jafari, Y.; Shirazi, J.K. Sustainability Spillover Effects of Social, Environment and Economy: Mapping Global Sustainable Development in a Systematic Analysis. *Asia-Pacific Journal of Regional Science* **2023**, *7*, 329–353, doi:10.1007/S41685-022-00231-0/METRICS.
166. Chen, C.W. Guidance on the Conceptual Design of Sustainable Product–Service Systems. *Sustainability* **2018**, *Vol. 10*, Page 2452 **2018**, *10*, 2452, doi:10.3390/SU10072452.
167. Cvijanović, D.; Ignjatijević, S.; Tankosić, J.V.; Cvijanović, V. Do Local Food Products Contribute to Sustainable Economic Development? *Sustainability* **2020**, *Vol. 12*, Page 2847 **2020**, *12*, 2847, doi:10.3390/SU12072847.
168. Tsakiridou, E.; Boutsouki, C.; Zotos, Y.; Mattas, K. Attitudes and Behaviour towards Organic Products: An Exploratory Study. *International Journal of Retail and Distribution Management* **2008**, *36*, 158–175, doi:10.1108/09590550810853093/FULL/XML.
169. Darnhofer, I.; Lindenthal, T.; Bartel-Kratochvil, R.; Zollitsch, W. Conventionalisation of Organic Farming Practices: From Structural Criteria towards an Assessment Based on Organic Principles. A Review. *Agronomy for Sustainable Development* **2010**, *30*, 67–81, doi:10.1051/AGRO/2009011.
170. Sogari, G.; Mora, C.; Menozzi, D. Factors Driving Sustainable Choice: The Case of Wine. *British Food Journal* **2016**, *118*, 632–646, doi:10.1108/BFJ-04-2015-0131/FULL/XML.
171. perspective, S.C.-O. agriculture: a global; 2006, undefined Organic Standards and Certification. *cabdirect.org* *S Courville Organic agriculture: a global perspective, 2006*•*cabdirect.org*.
172. Barrett, H.R.; Browne, A.W.; Harris, P.J.C.; Cadoret, K. Organic Certification and the UK Market: Organic Imports from Developing Countries. *Food Policy* **2002**, *27*, 301–318, doi:10.1016/S0306-9192(02)00036-2.
173. Moore, G. The Fair Trade Movement: Parameters, Issues and Future Research. *Journal of Business Ethics* **2004**, *53*, 73–86, doi:10.1023/B:BUSI.0000039400.57827.C3/METRICS.
174. Renard, M.C. Quality Certification, Regulation and Power in Fair Trade. *J Rural Stud* **2005**, *21*, 419–431, doi:10.1016/J.JRURSTUD.2005.09.002.
175. ZIAUL, I.Md.; SHUWEI, W. Environmental Sustainability: A Major Component of Sustainable Development. *International Journal of Environmental, Sustainability, and Social Science* **2023**, *4*, 900–907, doi:10.38142/IJESS.V4I2.296.
176. Getz, C.; Shreck, A. What Organic and Fair Trade Labels Do Not Tell Us: Towards a Place-Based Understanding of Certification. *Int J Consum Stud* **2006**, *30*, 490–501, doi:10.1111/J.1470-6431.2006.00533.X.
177. Hummen, T.; Sudheshwar, A. Fitness of Product and Service Design for Closed-Loop Material Recycling: A Framework and Indicator. *Resour Conserv Recycl* **2023**, *190*, 106661, doi:10.1016/J.RESCONREC.2022.106661.
178. Donatelli, A.; Casciaro, G.; Schioppa, M.; Palazzo, B.; Scalone, A.G.; Valentino, F.; Capodici, L.; De Pascalis, F.; Nacucchi, M.; Caretto, F. Thermoforming, Characterization, and in-House Recycling of Multi-Layered Laminated Composites Made with Polyamide 6/Recycled Carbon Fibers Hybrid Yarns-Based Plain Weave Fabrics. *J Compos Mater* **2023**, *57*, 3493–3513, doi:10.1177/00219983231187195/ASSET/IMAGES/LARGE/10.1177\_00219983231187195-IMG01.JPEG.
179. Hong, M.; Chen, E.Y.X. Chemically Recyclable Polymers: A Circular Economy Approach to Sustainability. *Green Chemistry* **2017**, *19*, 3692–3706, doi:10.1039/C7GC01496A.

180. Rusmevichientong, P.; Sumida, M.; Topaloglu, H. Dynamic Assortment Optimization for Reusable Products with Random Usage Durations. *https://doi.org/10.1287/mnsc.2019.3346* **2020**, *66*, 2820–2844, doi:10.1287/MNSC.2019.3346.
181. Sadeghi, A.H.; Bani, E.A.; Fallahi, A.; Handfield, R. Grey Wolf Optimizer and Whale Optimization Algorithm for Stochastic Inventory Management of Reusable Products in a Two-Level Supply Chain. *IEEE Access* **2023**, doi:10.1109/ACCESS.2023.3269292.
182. Dilanchiev, A.; Nuta, F.; Khan, I.; Khan, H. Urbanization, Renewable Energy Production, and Carbon Dioxide Emission in BSEC Member States: Implications for Climate Change Mitigation and Energy Markets. *Environmental Science and Pollution Research* **2023**, *30*, 67338–67350, doi:10.1007/S11356-023-27221-9/TABLES/7.
183. Zhang, Y.Q.; Li, L.; Sadiq, M.; Chien, F.S. The Impact of Non-Renewable Energy Production and Energy Usage on Carbon Emissions: Evidence from China. *https://doi.org/10.1177/0958305X221150432* **2023**, doi:10.1177/0958305X221150432.
184. Akroush, M.N.; Zuriekat, M.I.; Al Jabali, H.I.; Asfour, N.A. Determinants of Purchasing Intentions of Energy-Efficient Products: The Roles of Energy Awareness and Perceived Benefits. *International Journal of Energy Sector Management* **2019**, *13*, 128–148, doi:10.1108/IJESM-05-2018-0009/FULL/XML.
185. Ha, H.Y.; Janda, S. Predicting Consumer Intentions to Purchase Energy-Efficient Products. *Journal of Consumer Marketing* **2012**, *29*, 461–469, doi:10.1108/07363761211274974/FULL/XML.
186. Svanes, E.; Void, M.; Møller, H.; Pettersen, M.K.; Larsen, H.; Hanssen, O.J. Sustainable Packaging Design: A Holistic Methodology for Packaging Design. *Packaging Technology and Science* **2010**, *23*, 161–175, doi:10.1002/PTS.887.
187. Boz, Z.; Korhonen, V.; Sand, C.K. Consumer Considerations for the Implementation of Sustainable Packaging: A Review. *Sustainability* **2020**, *Vol. 12*, Page 2192 **2020**, *12*, 2192, doi:10.3390/SU12062192.
188. Beiras, R.; López-Ibáñez, S. A Practical Tool for the Assessment of Polymer Biodegradability in Marine Environments Guides the Development of Truly Biodegradable Plastics. *Polymers* **2023**, *Vol. 15*, Page 974 **2023**, *15*, 974, doi:10.3390/POLYM15040974.
189. Paul-Pont, I.; Ghiglione, J.F.; Gastaldi, E.; Ter Halle, A.; Huvet, A.; Bruzard, S.; Lagarde, F.; Galgani, F.; Duflos, G.; George, M.; et al. Discussion about Suitable Applications for Biodegradable Plastics Regarding Their Sources, Uses and End of Life. *Waste Management* **2023**, *157*, 242–248, doi:10.1016/J.WASMAN.2022.12.022.
190. Kang, D.; Jaisankar, R.; Murugesan, V.; Suvitha, K.; Narayanamoorthy, S.; Omar, A.H.; Arshad, N.I.; Ahmadian, A. A Novel MCDM Approach to Selecting a Biodegradable Dynamic Plastic Product: A Probabilistic Hesitant Fuzzy Set-Based COPRAS Method. *J Environ Manage* **2023**, *340*, 117967, doi:10.1016/J.JENVMAN.2023.117967.
191. Beattie, G.; Sale, L. Explicit and Implicit Attitudes to Low and High Carbon Footprint Products. *The International Journal of Environmental, Cultural, Economic, and Social Sustainability: Annual Review* **2009**, *5*, 191–206, doi:10.18848/1832-2077/CGP/V05I04/54652.
192. He, B.; Li, B.; Zhu, X. Carbon Footprint Prediction Method for Linkage Mechanism Design. *Environmental Science and Pollution Research* **2023**, *30*, 60150–60167, doi:10.1007/S11356-023-26556-7/METRICS.
193. Li, C.; Solangi, Y.A.; Ali, S. Evaluating the Factors of Green Finance to Achieve Carbon Peak and Carbon Neutrality Targets in China: A Delphi and Fuzzy AHP Approach. *Sustainability* **2023**, *Vol. 15*, Page 2721 **2023**, *15*, 2721, doi:10.3390/SU15032721.
194. Birkenberg, A.; Narjes, M.E.; Weinmann, B.; Birner, R. The Potential of Carbon Neutral Labeling to Engage Coffee Consumers in Climate Change Mitigation. *J Clean Prod* **2021**, *278*, 123621, doi:10.1016/J.JCLEPRO.2020.123621.
195. Mboga, J. ETHICAL SOURCING TO ENSURE SUSTAINABILITY IN THE FOOD AND BEVERAGE INDUSTRY AND ELICITING MILLENNIAL PERSPECTIVES. *European Journal of Economic and Financial Research* **2017**, *0*, doi:10.46827/EJEFR.V0I0.222.

196. Stamboulakis, D.; Sanderson, J. Certifying Biodiversity: The Union for Ethical BioTrade and the Search for Ethical Sourcing. *Journal of Environmental Law* **2020**, *32*, 503–528, doi:10.1093/JEL/EQAA013.
197. Andersson, S.; Svensson, G.; Otero-Neira, C.; Laurell, H.; Lindgren, J.; Karlsson, N.P.E. Sustainable Development Considerations in Supply Chains: Firms' Relationships with Stakeholders in Their Business Sustainability Practices—A Triangular Comparison. *Bus Strategy Environ* **2023**, *32*, 1885–1899, doi:10.1002/BSE.3225.
198. Goula, A.M.; Adamopoulos, K.G. A Method for Preparing a Novel Solid Product from Olive Mill Wastewater: Wastewater Characterization and Product Recovery. *Drying Technology* **2013**, *31*, 339–349, doi:10.1080/07373937.2012.736114.
199. Lyu, F.; Zhang, H.; Dang, C.; Gong, X. A Novel Framework for Water Accounting and Auditing for Efficient Management of Industrial Water Use. *J Clean Prod* **2023**, *395*, 136458, doi:10.1016/J.JCLEPRO.2023.136458.
200. Kwon, H.J.; Ahn, M.; Kang, J. The Effects of Knowledge Types on Consumer Decision Making for Non-Toxic Housing Materials and Products. *Sustainability* **2021**, *Vol. 13*, Page 11024 **2021**, *13*, 11024, doi:10.3390/SU131911024.
201. Agostini, V.O.; Macedo, A.J.; Muxagata, E.; da Silva, M.V.; Pinho, G.L.L. Natural and Non-Toxic Products from Fabaceae Brazilian Plants as a Replacement for Traditional Antifouling Biocides: An Inhibition Potential against Initial Biofouling. *Environmental Science and Pollution Research* **2019**, *26*, 27112–27127, doi:10.1007/S11356-019-05744-4/METRICS.
202. Sheehan, K.B.; Lee, J. What's Cruel About Cruelty Free: An Exploration of Consumers, Moral Heuristics, and Public Policy. *Journal of Animal Ethics* **2014**, *4*, 1–15, doi:10.5406/JANIMALETHICS.4.2.0001.
203. Grappe, C.G.; Lombart, C.; Louis, D.; Durif, F. "Not Tested on Animals": How Consumers React to Cruelty-Free Cosmetics Proposed by Manufacturers and Retailers? *International Journal of Retail and Distribution Management* **2021**, *49*, 1532–1553, doi:10.1108/IJRDM-12-2020-0489/FULL/XML.
204. Yu, S.; Lee, J. The Effects of Consumers' Perceived Values on Intention to Purchase Upcycled Products. *Sustainability* **2019**, *Vol. 11*, Page 1034 **2019**, *11*, 1034, doi:10.3390/SU11041034.
205. Adıgüzel, F.; Donato, C. Proud to Be Sustainable: Upcycled versus Recycled Luxury Products. *J Bus Res* **2021**, *130*, 137–146, doi:10.1016/J.JBUSRES.2021.03.033.
206. Jekanowski, M.D.; Williams, D.R.; Schiek, W.A. Consumers' Willingness to Purchase Locally Produced Agricultural Products: An Analysis of an Indiana Survey. *Agric Resour Econ Rev* **2000**, *29*, 43–53, doi:10.1017/S1068280500001428.
207. Wägeli, S.; Janssen, M.; Hamm, U. Organic Consumers' Preferences and Willingness-to-Pay for Locally Produced Animal Products. *Int J Consum Stud* **2016**, *40*, 357–367, doi:10.1111/IJCS.12262.
208. Benoit-Norris, C.; Cavan, D.A.; Norris, G. Identifying Social Impacts in Product Supply Chains: Overview and Application of the Social Hotspot Database. *Sustainability* **2012**, *Vol. 4*, Pages 1946-1965 **2012**, *4*, 1946–1965, doi:10.3390/SU4091946.
209. Norris, G.A. Social Impacts in Product Life Cycles: Towards Life Cycle Attribute Assessment. *International Journal of Life Cycle Assessment* **2006**, *11*, 97–104, doi:10.1065/LCA2006.04.017/METRICS.
210. Schieweck, A.; Bock, M.C. Emissions from Low-VOC and Zero-VOC Paints – Valuable Alternatives to Conventional Formulations Also for Use in Sensitive Environments? *Build Environ* **2015**, *85*, 243–252, doi:10.1016/J.BUILDENV.2014.12.001.
211. Silva, G.V.; Martins, A.O.; Martins, S.D.S.; Mata, T.M. Low-VOC Emission Label Proposal for Facemask Safety Based on Respiratory and Skin Health Criteria. *Environments - MDPI* **2023**, *10*, 10, doi:10.3390/ENVIRONMENTS10010010/S1.
212. GitHub - Rkonstadinou/Gpt-Sustainable-Products-Chrome-App: A GPT-Powered Software for Identifying Sustainable Product Features Available online:

- <https://github.com/rkonstadinos/gpt-sustainable-products-chrome-app> (accessed on 14 November 2023).
213. Pandas - Python Data Analysis Library Available online: <https://pandas.pydata.org/> (accessed on 14 November 2023).
  214. Chrome Extensions Getting Started Guides - Chrome for Developers Available online: <https://developer.chrome.com/docs/extensions/mv3/getstarted/> (accessed on 14 November 2023).
  215. Host, Run, and Code Python in the Cloud: PythonAnywhere Available online: <https://www.pythonanywhere.com/> (accessed on 14 November 2023).
  216. Anaconda | Anaconda Acquires PythonAnywhere to Increase Python... Available online: <https://www.anaconda.com/blog/anaconda-acquires-pythonanywhere> (accessed on 14 November 2023).
  217. Welcome to Flask – Flask Documentation (2.3.x) Available online: <https://flask.palletsprojects.com/en/2.3.x/> (accessed on 14 November 2023).
  218. Models - OpenAI API Available online: <https://platform.openai.com/docs/models/overview> (accessed on 14 November 2023).
  219. OpenAI Help Center Available online: <https://help.openai.com/en/articles/6779149-how-do-text-davinci-002-and-text-davinci-003-differ> (accessed on 14 November 2023).
  220. GPT-4 Available online: <https://openai.com/gpt-4> (accessed on 14 November 2023).
  221. Ali, M.N.Y.; Rahman, M.L.; Chaki, J.; Dey, N.; Santosh, K.C. Machine Translation Using Deep Learning for Universal Networking Language Based on Their Structure. *International Journal of Machine Learning and Cybernetics* **2021**, *12*, 2365–2376, doi:10.1007/S13042-021-01317-5/METRICS.
  222. Yılmaz Benk, G.; Badur, B.; Mardikyan, S. A New 360° Framework to Predict Customer Lifetime Value for Multi-Category E-Commerce Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence. *Information* **2022**, *13*, 373, doi:10.3390/INFO13080373.
  223. De Caigny, A.; Coussement, K.; Verbeke, W.; Idbenjra, K.; Phan, M. Uplift Modeling and Its Implications for B2B Customer Churn Prediction: A Segmentation-Based Modeling Approach. *Industrial Marketing Management* **2021**, *99*, 28–39, doi:10.1016/J.INDMARMAN.2021.10.001.
  224. Jain, H.; Khunteta, A.; Srivastava, S. Churn Prediction in Telecommunication Using Logistic Regression and Logit Boost. *Procedia Comput Sci* **2020**, *167*, 101–112, doi:10.1016/J.PROCS.2020.03.187.
  225. Khattak, A.; Mehak, Z.; Ahmad, H.; Asghar, M.U.; Asghar, M.Z.; Khan, A. Customer Churn Prediction Using Composite Deep Learning Technique. *Scientific Reports* **2023**, *13*, 1–17, doi:10.1038/s41598-023-44396-w.
  226. J, S.; Gangadhar, C.; Arora, R.K.; Renjith, P.N.; Bamini, J.; Chincholkar, Y. devidas E-Commerce Customer Churn Prevention Using Machine Learning-Based Business Intelligence Strategy. *Measurement: Sensors* **2023**, *27*, 100728, doi:10.1016/J.MEASEN.2023.100728.
  227. Rosário, A.T.; Dias, J.C. How Has Data-Driven Marketing Evolved: Challenges and Opportunities with Emerging Technologies. *International Journal of Information Management Data Insights* **2023**, *3*, 100203, doi:10.1016/J.JJIMEI.2023.100203.
  228. Ahn, J.; Hwang, J.; Kim, D.; Choi, H.; Kang, S. A Survey on Churn Analysis in Various Business Domains. *IEEE Access* **2020**, *8*, 220816–220839, doi:10.1109/ACCESS.2020.3042657.
  229. Larsson, A.; Broström, E. Ensuring Customer Retention: Insurers' Perception of Customer Loyalty. *Marketing Intelligence and Planning* **2020**, *38*, 151–166, doi:10.1108/MIP-02-2019-0106/FULL/PDF.
  230. Christy, A.J.; Umamakeswari, A.; Priyatharsini, L.; Neyaa, A. RFM Ranking – An Effective Approach to Customer Segmentation. *Journal of King Saud University - Computer and Information Sciences* **2021**, *33*, 1251–1257, doi:10.1016/J.JKSUCI.2018.09.004.
  231. Pustokhina, I. V.; Pustokhin, D.A.; RH, A.; Jayasankar, T.; Jeyalakshmi, C.; Díaz, V.G.; Shankar, K. Dynamic Customer Churn Prediction Strategy for Business Intelligence Using Text Analytics

- with Evolutionary Optimization Algorithms. *Inf Process Manag* **2021**, 58, 102706, doi:10.1016/J.IPM.2021.102706.
232. GitHub - Kroumeliotis/Churn-Modeling-in-E-Commerce: Churn Modeling in E-Commerce Available online: <https://github.com/kroumeliotis/Churn-Modeling-in-E-commerce> (accessed on 12 February 2024).
  233. Anderson, S.; Bohman, P.R.; Burmeister, O.K.; Sampson-Wild, G. User Needs and E-Government Accessibility: The Future Impact of WCAG 2.0. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2004**, 3196, 289–304, doi:10.1007/978-3-540-30111-0\_25/COVER.
  234. Sheppard, D. Beginning Progressive Web App Development: Creating a Native App Experience on the Web. *Beginning Progressive Web App Development: Creating a Native App Experience on the Web* **2017**, 1–266, doi:10.1007/978-1-4842-3090-9/COVER.
  235. Rojas, C. Building Progressive Web Applications with Vue.js: Reliable, Fast, and Engaging Apps with Vue.js. *Building Progressive Web Applications with Vue.js: Reliable, Fast, and Engaging Apps with Vue.js* **2019**, 1–233, doi:10.1007/978-1-4842-5334-2/COVER.
  236. Shimray, S.R.; Ramaiah, C.K. Use of Internet through Mobile Devices: A Survey. *Journal of Information and Knowledge* **2019**, 100–105, doi:10.17821/SRELS/2019/V56I2/141631.
  237. Mobile Internet & Apps | Statista Available online: <https://www.statista.com/markets/424/topic/538/mobile-internet-apps/#overview> (accessed on 9 November 2023).
  238. Malavolta, I.; Procaccianti, G.; Noorland, P.; Vukmirovic, P. Assessing the Impact of Service Workers on the Energy Efficiency of Progressive Web Apps. *Proceedings - 2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems, MOBILESoft 2017* **2017**, 35–45, doi:10.1109/MOBILESOFT.2017.7.
  239. Madsen, M.; Lhoták, O.; Tip, F. A Model for Reasoning about JavaScript Promises. *Proceedings of the ACM on Programming Languages* **2017**, 1, doi:10.1145/3133910.
  240. Chinprutthiwong, P.; Vardhan, R.; Yang, G.L.; Gu, G. Security Study of Service Worker Cross-Site Scripting. *ACM International Conference Proceeding Series* **2020**, 12, 643–654, doi:10.1145/3427228.3427290.
  241. ServiceWorkerContainer: Register() Method - Web APIs | MDN Available online: <https://developer.mozilla.org/en-US/docs/Web/API/ServiceWorkerContainer/register> (accessed on 9 November 2023).
  242. Is Service Worker Ready? Available online: <https://jakearchibald.github.io/isserviceworkerready/> (accessed on 9 November 2023).
  243. Gambhir, A.; Raj, G. Analysis of Cache in Service Worker and Performance Scoring of Progressive Web Application. *Proceedings on 2018 International Conference on Advances in Computing and Communication Engineering, ICACCE 2018* **2018**, 294–299, doi:10.1109/ICACCE.2018.8441715.
  244. (PDF) Dawning of Progressive Web Applications (PWA): Edging Out the Pitfalls of Traditional Mobile Development Available online: [https://www.researchgate.net/publication/343472764\\_Dawning\\_of\\_Progressive\\_Web\\_Applications\\_PWA\\_Edging\\_Out\\_the\\_Pitfalls\\_of\\_Traditional\\_Mobile\\_Development](https://www.researchgate.net/publication/343472764_Dawning_of_Progressive_Web_Applications_PWA_Edging_Out_the_Pitfalls_of_Traditional_Mobile_Development) (accessed on 9 November 2023).
  245. Architecture | Web.Dev Available online: <https://web.dev/learn/pwa/architecture/#what> (accessed on 9 November 2023).
  246. Web.Dev Available online: <https://web.dev/case-studies> (accessed on 9 November 2023).
  247. Web Accessibility | Shaping Europe’s Digital Future Available online: <https://digital-strategy.ec.europa.eu/en/policies/web-accessibility> (accessed on 9 November 2023).
  248. Abuaddous, Hayfa.Y.; Jali, M.Z.; Basir, N. Web Accessibility Challenges. *International Journal of Advanced Computer Science and Applications* **2016**, 7, doi:10.14569/IJACSA.2016.071023.

249. (PDF) Present and Future of Web Content Accessibility: An Analysis Available online: [https://www.researchgate.net/publication/289564468\\_Present\\_and\\_future\\_of\\_web\\_content\\_accessibility\\_An\\_analysis](https://www.researchgate.net/publication/289564468_Present_and_future_of_web_content_accessibility_An_analysis) (accessed on 9 November 2023).
250. What Is the Americans with Disabilities Act (ADA)? | ADA National Network Available online: <https://adata.org/learn-about-ada> (accessed on 9 November 2023).
251. Web Content Accessibility Guidelines (WCAG) 2.1 Available online: <https://www.w3.org/TR/WCAG21/> (accessed on 9 November 2023).
252. Sohaib, O.; Hussain, W.; Khalid Badini, M. User Experience (UX) and the Web Accessibility Standards. *IJCSI International Journal of Computer Science Issues* **2011**, *8*.
253. ISO 9241-151:2008 - Ergonomics of Human-System Interaction — Part 151: Guidance on World Wide Web User Interfaces Available online: <https://www.iso.org/standard/37031.html> (accessed on 9 November 2023).
254. Çağiltay, K.; Alacam, O.; Ocak, N.; Erdal, F. Developing ISO 9241-151 Product Certification Process: Challenges. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2013**, *8015 LNCS*, 334–341, doi:10.1007/978-3-642-39253-5\_36/COVER.
255. IT Accessibility Laws and Policies | Section508.Gov Available online: <https://www.section508.gov/manage/laws-and-policies/> (accessed on 9 November 2023).
256. Campoverde-Molina, M.; Lujan-Mora, S.; Garcia, L.V. Empirical Studies on Web Accessibility of Educational Websites: A Systematic Literature Review. *IEEE Access* **2020**, *8*, 91676–91700, doi:10.1109/ACCESS.2020.2994288.
257. Web Content Accessibility Guidelines 1.0 Available online: <https://www.w3.org/TR/WAI-WEBCONTENT/> (accessed on 9 November 2023).
258. Rodríguez, G.; Pérez, J.; Cueva, S.; Torres, R. A Framework for Improving Web Accessibility and Usability of Open Course Ware Sites. *Comput Educ* **2017**, *109*, 197–215, doi:10.1016/J.COMPEDU.2017.02.013.
259. Understanding WCAG 2.2 | WAI | W3C Available online: <https://www.w3.org/WAI/WCAG22/Understanding/> (accessed on 9 November 2023).
260. McEwen, B.S.; Sapolsky, R.M. Stress and Cognitive Function. *Curr Opin Neurobiol* **1995**, *5*, 205–216, doi:10.1016/0959-4388(95)80028-X.
261. Aizpurua, A.; Arrue, M.; Vigo, M. Prejudices, Memories, Expectations and Confidence Influence Experienced Accessibility on the Web. *Comput Human Behav* **2015**, *51*, 152–160, doi:10.1016/J.CHB.2015.04.035.
262. Brajnik, G.; Yesilada, Y.; Harper, S. Is Accessibility Conformance an Elusive Property? A Study of Validity and Reliability of WCAG 2.0. *ACM Transactions on Accessible Computing (TACCESS)* **2012**, *4*, doi:10.1145/2141943.2141946.
263. Karreman, J.; Van Der Geest, T.; Buursink, E. Accessible Website Content Guidelines for Users with Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities* **2007**, *20*, 510–518, doi:10.1111/J.1468-3148.2006.00353.X.
264. What Makes a Good Progressive Web App? | Articles | Web.Dev Available online: <https://web.dev/articles/pwa-checklist> (accessed on 9 November 2023).
265. Pimentel, J.F.; Murta, L.; Braganholo, V.; Freire, J. A Large-Scale Study about Quality and Reproducibility of Jupyter Notebooks. *IEEE International Working Conference on Mining Software Repositories* **2019**, *2019-May*, 507–517, doi:10.1109/MSR.2019.00077.
266. Vollenwyder, B.; Iten, G.H.; Brühlmann, F.; Opwis, K.; Mekler, E.D. Salient Beliefs Influencing the Intention to Consider Web Accessibility. *Comput Human Behav* **2019**, *92*, 352–360, doi:10.1016/J.CHB.2018.11.016.
267. Gandon, F.; Hall, W. A Never-Ending Project for Humanity Called “the Web.” *WWW 2022 - Proceedings of the ACM Web Conference 2022* **2022**, 3480–3487, doi:10.1145/3485447.3514195.
268. Heuer, J.; Hund, J.; Pfaff, O. Toward the Web of Things: Applying Web Technologies to the Physical World. *Computer (Long Beach Calif)* **2015**, *48*, 34–42, doi:10.1109/MC.2015.152.

269. Total Number of Websites - Internet Live Stats Available online: <https://www.internetlivestats.com/total-number-of-websites/> (accessed on 7 November 2023).
270. SEReleC (Search Engine Result Refinement and Classification) - a Meta Search Engine Based on Combinatorial Search and Search Keyword Based Link Classification | IEEE Conference Publication | IEEE Xplore Available online: <https://ieeexplore.ieee.org/document/6215917> (accessed on 8 November 2023).
271. Google Search Essentials (Formerly Webmaster Guidelines) | Google Search Central | Documentation | Google for Developers Available online: <https://developers.google.com/search/docs/essentials> (accessed on 8 November 2023).
272. Luh, C.J.; Yang, S.A.; Huang, T.L.D. Estimating Google's Search Engine Ranking Function from a Search Engine Optimization Perspective. *Online Information Review* **2016**, *40*, 239–255, doi:10.1108/OIR-04-2015-0112/FULL/XML.
273. Patil, V.M.; Patil, A. V. SEO: On-Page + Off-Page Analysis. *2018 International Conference on Information, Communication, Engineering and Technology, ICICET 2018* **2018**, doi:10.1109/ICICET.2018.8533836.
274. Matošević, G.; Dobša, J.; Mladenčić, D. Using Machine Learning for Web Page Classification in Search Engine Optimization. *Future Internet* **2021**, *Vol. 13*, Page 9 **2021**, *13*, 9, doi:10.3390/FI13010009.
275. Iqbal, M.; Khalid, M.N.; Manzoor, D.A.; Abid, M.M.; Shaikh, N.A. Search Engine Optimization (SEO): A Study of Important Key Factors in Achieving a Better Search Engine Result Page (SERP) Position. *Sukkur IBA Journal of Computing and Mathematical Sciences* **2022**, *6*, 1–15, doi:10.30537/SJCMS.V6I1.924.
276. Zhou, H.; Qin, S.; Liu, J.; Chen, J. Study on Website Search Engine Optimization. *Proceedings - 2012 International Conference on Computer Science and Service System, CSSS 2012* **2012**, 930–933, doi:10.1109/CSSS.2012.236.
277. Gudivada, V.N.; Rao, D.; Paris, J. Understanding Search-Engine Optimization. *Computer (Long Beach Calif)* **2015**, *48*, 43–52, doi:10.1109/MC.2015.297.
278. Mo, Y. A Study on Tactics for Corporate Website Development Aiming at Search Engine Optimization. *2nd International Workshop on Education Technology and Computer Science, ETCS 2010* **2010**, *3*, 673–675, doi:10.1109/ETCS.2010.230.
279. Li, K.; Lin, M.; Lin, Z.; Xing, B. Running and Chasing - The Competition between Paid Search Marketing and Search Engine Optimization. *Proceedings of the Annual Hawaii International Conference on System Sciences* **2014**, 3110–3119, doi:10.1109/HICSS.2014.640.
280. Lemos, J.Y.; Joshi, A.R. Search Engine Optimization to Enhance User Interaction. *Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017* **2017**, 398–402, doi:10.1109/I-SMAC.2017.8058379.
281. Yalçın, N.; Köse, U. What Is Search Engine Optimization: SEO? *Procedia Soc Behav Sci* **2010**, *9*, 487–493, doi:10.1016/J.SBSPRO.2010.12.185.
282. Krrabaj, S.; Baxhaku, F.; Sadrijaj, D. Investigating Search Engine Optimization Techniques for Effective Ranking: A Case Study of an Educational Site. *2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings* **2017**, doi:10.1109/MECO.2017.7977137.
283. Chhabra, S.; Mittal, R.; Sarkar, D. Inducing Factors for Search Engine Optimization Techniques: A Comparative Analysis. *India International Conference on Information Processing, IICIP 2016 - Proceedings* **2017**, doi:10.1109/IICIP.2016.7975341.
284. Invernizzi, L.; Comparetti, P.M.; Benvenuti, S.; Kruegel, C.; Cova, M.; Vigna, G. EvilSeed: A Guided Approach to Finding Malicious Web Pages. *Proc IEEE Symp Secur Priv* **2012**, 428–442, doi:10.1109/SP.2012.33.
285. Zhang, Q.; Wang, D.Y.; Voelker, G.M. DSpin: Detecting Automatically Spun Content on the Web. **2014**, doi:10.14722/NDSS.2014.23004.

286. Invernizzi, L.; Thomas, K.; Kapravelos, A.; Comanescu, O.; Picod, J.M.; Bursztein, E. Cloak of Visibility: Detecting When Machines Browse a Different Web. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016* **2016**, 743–758, doi:10.1109/SP.2016.50.
287. Savoska, S.; Naka, N.; Manevska, V.; Ristevski, B. Search Engine Optimization on PHP Based Web Pages in Practice. *Journal of Emerging research and solutions in ICT* **2016**, *1*, 45–58, doi:10.20544/ERSICT.01.16.P05.
288. Killoran, J.B. How to Use Search Engine Optimization Techniques to Increase Website Visibility. *IEEE Trans Prof Commun* **2013**, *56*, 50–66, doi:10.1109/TPC.2012.2237255.
289. Kohli, S.; Kaur, S.; Singh, G. A Website Content Analysis Approach Based on Keyword Similarity Analysis. *Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012* **2012**, 254–257, doi:10.1109/WI-IAT.2012.212.
290. Szymanski, G.; Lipinski, P. Model of the Effectiveness of Google Adwords Advertising Activities. *International Scientific and Technical Conference on Computer Sciences and Information Technologies* **2018**, *2*, 98–101, doi:10.1109/STC-CSIT.2018.8526633.
291. Kumar, G.; Paul, R.K. Literature Review on On-Page & Off-Page SEO for Ranking Purpose. *United International Journal for Research & Technology (UIJRT) United International Journal for Research & Technology (UIJRT)* **2020**, *1*, 30–34.
292. Wang, F.; Li, Y.; Zhang, Y. An Empirical Study on the Search Engine Optimization Technique and Its Outcomes. *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce, AIMSEC 2011 - Proceedings* **2011**, 2767–2770, doi:10.1109/AIMSEC.2011.6011361.
293. What Are Title Tags? [Plus FREE Meta Title Preview Tool] - Moz Available online: <https://moz.com/learn/seo/title-tag> (accessed on 8 November 2023).
294. Zuze, H.; Weideman, M. Keyword Stuffing and the Big Three Search Engines. *Online Information Review* **2013**, *37*, 268–286, doi:10.1108/OIR-11-2011-0193/FULL/XML.
295. SEO Starter Guide: The Basics | Google Search Central | Documentation | Google for Developers Available online: <https://support.google.com/webmasters/answer/7451184> (accessed on 8 November 2023).
296. Optimize Your Site for Search Engines (for Beginners) - Search Console Help Available online: [https://support.google.com/webmasters/answer/9128678?visit\\_id=638350360443139811-2218532793&rd=1](https://support.google.com/webmasters/answer/9128678?visit_id=638350360443139811-2218532793&rd=1) (accessed on 8 November 2023).
297. Kayes, I.; Rahma Akhter, S. Impact of Copywriting in Marketing Communication. **2018**.
298. Swati, P.P.; Ajay, P.S. Search Engine Optimization: A Study. *Research Journal of Computer and Information Technology Sciences* **2013**, *1*, 10–13.
299. Dragić, R.; Kaurin, T. Significance and Impact of Meta Tags on Search Engine Results Pages. *INFOTEH-JAHORINA* **2012**, *11*, 751.
300. Kumar Gunjan, V.; Kumari, M.; Kumar, A.; appa rao, A. Search Engine Optimization with Google. *IJCSI International Journal of Computer Science Issues* **2012**, *9*, 206–214.
301. Minh, D.P.; Van, T.L.; Le Dinh, T. Identification of Paths and Parameters in RESTful URLs for the Detection of Web Attacks. *2017 4th NAFOSTED Conference on Information and Computer Science, NICS 2017 - Proceedings* **2017**, 2017-January, 110–115, doi:10.1109/NAFOSTED.2017.8108048.
302. Zhu, C.; Wu, G. Research and Analysis of Search Engine Optimization Factors Based on Reverse Engineering. *Proceedings - 3rd International Conference on Multimedia Information Networking and Security, MINES 2011* **2011**, 225–228, doi:10.1109/MINES.2011.99.
303. Nen, M.; Popa, V.; Scurtu, A.; Unc, R.L.; Nen, M.; Popa, V.; Scurtu, A.; Unc, R.L. The Computer Management – SEO Audit. *REVISTA DE MANAGEMENT COMPARAT INTERNATIONAL/REVIEW OF INTERNATIONAL COMPARATIVE MANAGEMENT* **2017**, *18*, 297–307.
304. Patzer, A.; Moodie, M. The Front Controller Pattern. *Foundations of JSP Design Patterns* **2004**, 143–165, doi:10.1007/978-1-4302-0736-8\_7.

305. Zhang, S.; Cabage, N. Does SEO Matter? Increasing Classroom Blog Visibility through Search Engine Optimization. *Proceedings of the Annual Hawaii International Conference on System Sciences* **2013**, 1610–1619, doi:10.1109/HICSS.2013.184.
306. Should I Change My URLs for SEO? - Moz Available online: <https://moz.com/blog/should-i-change-my-urls-for-seo> (accessed on 8 November 2023).
307. What Are Meta Descriptions And How to Write Them - Moz Available online: <https://moz.com/learn/seo/meta-description> (accessed on 8 November 2023).
308. W3C Standards and Drafts | W3C Available online: <https://www.w3.org/TR/> (accessed on 8 November 2023).
309. Rovira, C.; Codina, L.; Lopezosa, C. Language Bias in the Google Scholar Ranking Algorithm. *Future Internet* **2021**, Vol. 13, Page 31 **2021**, 13, 31, doi:10.3390/FI13020031.
310. Here's What You Need to Know about Image Optimization for SEO Available online: <https://searchengineland.com/heres-what-you-need-to-know-about-image-optimization-for-seo-316046> (accessed on 8 November 2023).
311. URLs with "Image Size Is over 100 KB" [Guide on How to Fix Them] | Sitechecker Available online: <https://sitechecker.pro/site-audit-issues/image-size-100-kb/> (accessed on 8 November 2023).
312. Wang, H.; Tang, G.; Wu, K.; Fan, J. Speeding up Multi-CDN Content Delivery via Traffic Demand Reshaping. *Proc Int Conf Distrib Comput Syst* **2018**, 2018-July, 422–433, doi:10.1109/ICDCS.2018.00049.
313. Bigham, J.P. Increasing Web Accessibility by Automatically Judging Alternative Text Quality. *International Conference on Intelligent User Interfaces, Proceedings IUI* **2007**, 349–352, doi:10.1145/1216295.1216364.
314. H37: Using Alt Attributes on Img Elements | Techniques for WCAG 2.0 Available online: <https://www.w3.org/TR/WCAG20-TECHS/H37.html> (accessed on 8 November 2023).
315. WebAIM: Alternative Text Available online: <https://webaim.org/techniques/alttext/> (accessed on 8 November 2023).
316. Bahtar, A.Z.; Muda, M. The Impact of User – Generated Content (UGC) on Product Reviews towards Online Purchasing – A Conceptual Framework. *Procedia Economics and Finance* **2016**, 37, 337–342, doi:10.1016/S2212-5671(16)30134-4.
317. Amblee, N.; Bui, T. Harnessing the Influence of Social Proof in Online Shopping: The Effect of Electronic Word of Mouth on Sales of Digital Microproducts. *International Journal of Electronic Commerce* **2011**, 16, 91–113, doi:10.2753/JEC1086-4415160205.
318. Gabir, H.H.; Karrar, A.Z. The Effect of Website's Design Factors on Conversion Rate in E-Commerce. *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering, ICCCEE 2018* **2018**, doi:10.1109/ICCCEE.2018.8515842.
319. Schema.Org - Schema.Org Available online: <https://schema.org/> (accessed on 8 November 2023).
320. Navarrete, R.; Lujan-Mora, S. Microdata with Schema Vocabulary: Improvement Search Results Visualization of Open Educational Resources. *Iberian Conference on Information Systems and Technologies, CISTI* **2018**, 2018-June, 1–6, doi:10.23919/CISTI.2018.8399222.
321. RDFa 1.1 Primer - Third Edition Available online: <https://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/> (accessed on 8 November 2023).
322. Lanthaler, M.; Gütl, C. On Using JSON-LD to Create Evolvable RESTful Services. *ACM International Conference Proceeding Series* **2012**, 25–32, doi:10.1145/2307819.2307827.
323. RDF AND JSON-LD UseCases - Data on the Web Best Practices Available online: [https://www.w3.org/2013/dwbp/wiki/RDF\\_AND\\_JSON-LD\\_UseCases](https://www.w3.org/2013/dwbp/wiki/RDF_AND_JSON-LD_UseCases) (accessed on 8 November 2023).
324. Design and Development of Automated Tool to Study Sitemap as Design Issue in Websites | IEEE Conference Publication | IEEE Xplore Available online: <https://ieeexplore.ieee.org/document/7100303> (accessed on 8 November 2023).
325. Hammersley, B. Content Syndication with RSS. **2003**.

326. James Lewin An Introduction to RSS News Feeds Available online: <https://xml.coverpages.org/lewin-w-rss.pdf> (accessed on 8 November 2023).
327. Ma, D. Offering RSS Feeds: Does It Help to Gain Competitive Advantage? *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS 2009*, doi:10.1109/HICSS.2009.327.
328. Ragavan, N. Efficient Key Hash Indexing Scheme with Page Rank for Category Based Search Engine Big Data. *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017 2018*, 2018-February, 1–6, doi:10.1109/ITCOSP.2017.8303118.
329. Robots.Txt File - Joomla! Documentation Available online: [https://docs.joomla.org/Robots.txt\\_file](https://docs.joomla.org/Robots.txt_file) (accessed on 8 November 2023).
330. Youngblood, N.E.; MacKiewicz, J. A Usability Analysis of Municipal Government Website Home Pages in Alabama. *Gov Inf Q* **2012**, *29*, 582–588, doi:10.1016/J.GIQ.2011.12.010.
331. How To Add Breadcrumb (BreadcrumbList) Markup | Google Search Central | Documentation | Google for Developers Available online: <https://developers.google.com/search/docs/appearance/structured-data/breadcrumb> (accessed on 8 November 2023).
332. Search Engine Optimization: A Game of Page Ranking | IEEE Conference Publication | IEEE Xplore Available online: <https://ieeexplore.ieee.org/abstract/document/7100246> (accessed on 8 November 2023).
333. Sharma, S.; Verma, S. Optimizing Website Effectiveness Using Various SEO Techniques. *2020 7th International Conference on Signal Processing and Integrated Networks, SPIN 2020 2020*, 918–922, doi:10.1109/SPIN48934.2020.9070893.
334. Mobile-Friendly Test - Search Console Help Available online: <https://support.google.com/webmasters/answer/6352293?hl=en> (accessed on 9 November 2023).
335. Perakakis, E.; Ghinea, G. Smart Enough for the Web? A Responsive Web Design Approach to Enhancing the User Web Browsing Experience on Smart TVs. *IEEE Trans Hum Mach Syst* **2017**, *47*, 860–872, doi:10.1109/THMS.2017.2726821.
336. Hua, Z.; Xie, X.; Liu, H.; Lu, H.; Ma, W.Y. Design and Performance Studies of an Adaptive Scheme for Serving Dynamic Web Content in a Mobile Computing Environment. *IEEE Trans Mob Comput* **2006**, *5*, 1650–1661, doi:10.1109/TMC.2006.182.
337. Alnawaj'ha, F.Y.; Abutaha, M.S. Responsive Web Design Commitment by the Web Developers in Palestine. *2018 4th International Conference on Computer and Technology Applications, ICCTA 2018 2018*, 69–73, doi:10.1109/CATA.2018.8398658.
338. Mobile-Friendly Test - Google Search Console Available online: <https://search.google.com/test/mobile-friendly> (accessed on 9 November 2023).
339. Hoque, M.; Alsadoon, A.; Maag, A.; Prasad, P.W.C.; Elchouemi, A. Comprehensive Search Engine Optimization Model for Commercial Websites: Surgeon's Website in Sydney. *Journal of Software* **2018**, *12*, 43–56, doi:10.17706/JSW.13.1.43-56.
340. Kaur, S.; Kaur, K.; Kaur, P. An Empirical Performance Evaluation of Universities Website. *Int J Comput Appl* **2016**, *146*, 975–8887.
341. Lighthouse Overview - Chrome for Developers Available online: <https://developer.chrome.com/docs/lighthouse/overview/> (accessed on 9 November 2023).
342. Website Speed Test | Pingdom Tools Available online: <https://tools.pingdom.com/> (accessed on 9 November 2023).
343. HTTPS as a Ranking Signal | Google Search Central Blog | Google for Developers Available online: <https://developers.google.com/search/blog/2014/08/https-as-ranking-signal> (accessed on 9 November 2023).
344. Chrome Releases: Stable Channel Update for Desktop Available online: <https://chromereleases.googleblog.com/2018/07/stable-channel-update-for-desktop.html> (accessed on 9 November 2023).

345. Guha, R. V.; Brickley, D.; MacBeth, S. Schema.Org: Evolution of Structured Data on the Web. *Queue* **2015**, *13*, doi:10.1145/2857274.2857276.
346. Navarrete, R.; Luján-Mora, S. Use of Embedded Markup for Semantic Annotations in E-Government and e-Education Websites. *2017 4th International Conference on eDemocracy and eGovernment, ICEDEG 2017* **2017**, 71–78, doi:10.1109/ICEDEG.2017.7962515.
347. Get Started - Amp.Dev Available online: <https://amp.dev/documentation/> (accessed on 9 November 2023).
348. Phokeer, A.; Chavula, J.; Johnson, D.; Densmore, M.; Tyson, G.; Sathiaselan, A.; Feamster, N. On the Potential of Google AMP to Promote Local Content in Developing Regions. *2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019* **2019**, 80–87, doi:10.1109/COMSNETS.2019.8711393.
349. AMP Test - Google Search Console Available online: <https://search.google.com/test/amp> (accessed on 9 November 2023).
350. Gzip - GNU Project - Free Software Foundation Available online: <https://www.gnu.org/software/gzip/> (accessed on 9 November 2023).
351. Shroff, P.H.; Chaudhary, S.R. Critical Rendering Path Optimizations to Reduce the Web Page Loading Time. *2017 2nd International Conference for Convergence in Technology, I2CT 2017* **2017**, *2017-January*, 937–940, doi:10.1109/I2CT.2017.8226266.
352. Cache - Documentation - Twig - The Flexible, Fast, and Secure PHP Template Engine Available online: <https://twig.symfony.com/doc/3.x/tags/cache.html> (accessed on 9 November 2023).
353. Mod\_expires - Apache HTTP Server Version 2.4 Available online: [https://httpd.apache.org/docs/current/mod/mod\\_expires.html](https://httpd.apache.org/docs/current/mod/mod_expires.html) (accessed on 9 November 2023).
354. The Open Graph Protocol Available online: <https://ogp.me/> (accessed on 9 November 2023).
355. Free Plan Overview | Cloudflare Available online: <https://www.cloudflare.com/en-gb/plans/free/> (accessed on 9 November 2023).
356. Using Page Speed in Mobile Search Ranking | Google Search Central Blog | Google for Developers Available online: <https://developers.google.com/search/blog/2018/01/using-page-speed-in-mobile-search> (accessed on 24 November 2023).
357. Using Site Speed in Web Search Ranking | Google Search Central Blog | Google for Developers Available online: <https://developers.google.com/search/blog/2010/04/using-site-speed-in-web-search-ranking> (accessed on 9 November 2023).
358. White Hat SEO: What It Is & Techniques to Use Available online: <https://www.semrush.com/blog/white-hat-seo/> (accessed on 9 November 2023).
359. Ma, X. Research on Black Hat SEO Behaviour Measurement. *Proceedings of 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2018* **2018**, 1041–1045, doi:10.1109/IAEAC.2018.8577831.
360. Bello, R.-W.; Ootobo, F.N. Conversion of Website Users to Customers-The Black Hat SEO Technique. *International Journal of Advanced Research in Computer Science and Software Engineering* **2018**, *8*, 29, doi:10.23956/IJARCSSE.V8I6.714.
361. Van Goethem, T.; Miramirkhani, N.; Joosen, W.; Nikiforakis, N. Purchased Fame: Exploring the Ecosystem of Private Blog Networks. *AsiaCCS 2019 - Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* **2019**, 366–378, doi:10.1145/3321705.3329830.
362. Saura, J.R.; Reyes-Menendez, A.; Van Nostrand, C. Does SEO Matter for Startups? Identifying Insights from UGC Twitter Communities. *Informatics 2020, Vol. 7, Page 47* **2020**, *7*, 47, doi:10.3390/INFORMATICS7040047.
363. Damaševičius, R.; Suri, B.; Roumeliotis, K.I.; Tselikas, N.D. A Machine Learning Python-Based Search Engine Optimization Audit Software. *Informatics 2023, Vol. 10, Page 68* **2023**, *10*, 68, doi:10.3390/INFORMATICS10030068.
364. GitHub - Rkonstadinou/Python-Based-Seo-Audit-Tool: Python-Based SEO Audit Tool Available online: <https://github.com/rkonstadinou/python-based-seo-audit-tool> (accessed on 24 November 2023).

365. Free-Proxy · PyPI Available online: <https://pypi.org/project/free-proxy/> (accessed on 24 November 2023).
366. Effortlessly Scrape Accurate Search Results with Zenserp API Available online: <https://zenserp.com/> (accessed on 24 November 2023).
367. Moz - Moz API | SEO API Tool Available online: <https://moz.com/products/api> (accessed on 9 November 2023).
368. Science, W.M.-P. of the 9th P. in; 2010, undefined Data Structures for Statistical Computing in Python. *conference.scipy.org* W McKinney *Proceedings of the 9th Python in Science Conference, 2010* • *conference.scipy.org* **2010**.
369. Pedregosa FABIANPEDREGOSA, F.; Michel, V.; Grisel OLIVIERGRISEL, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *jmlr.org* F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel *the Journal of machine Learning research, 2011* • *jmlr.org* **2011**, 12, 2825–2830.
370. Joblib: Running Python Functions as Pipeline Jobs — Joblib 1.4.Dev0 Documentation Available online: <https://joblib.readthedocs.io/en/latest/> (accessed on 24 November 2023).
371. SEO Tools, Software and Articles | SEO Site Checkup Available online: <https://seositecheckup.com/> (accessed on 24 November 2023).
372. SEO Checker | Test Your Website for Free with Seobility Available online: <https://www.seobility.net/en/seocheck/> (accessed on 24 November 2023).
373. Free Backlink Checker by Ahrefs: Check Backlinks to Any Site Available online: <https://ahrefs.com/backlink-checker> (accessed on 24 November 2023).
374. Soobhany, R.; Anwar, S.; Razzak, I.; Damaševičius, R.; Zailskait' E-Jakšt' E, L. Usability and Security Testing of Online Links: A Framework for Click-Through Rate Prediction Using Deep Learning. *Electronics* 2022, Vol. 11, Page 400 **2022**, 11, 400, doi:10.3390/ELECTRONICS11030400.
375. Seomized/Seomized.Csv at Main · Kroumeliotis/Seomized · GitHub Available online: <https://github.com/kroumeliotis/seomized/blob/main/seomized.csv> (accessed on 9 November 2023).
376. Amazon Alexa Available online: <https://alexa.com/> (accessed on 9 November 2023).
377. Seomized/Seomized at Main · Kroumeliotis/Seomized · GitHub Available online: <https://github.com/kroumeliotis/seomized/tree/main/seomized> (accessed on 9 November 2023).
378. Pagespeedapi: Runpagespeed | PageSpeed Insights | Google for Developers Available online: <https://developers.google.com/speed/docs/insights/v4/reference/pagespeedapi/runpagespeed> (accessed on 9 November 2023).
379. Seomized/Seomized.Sql at Main · Kroumeliotis/Seomized · GitHub Available online: <https://github.com/kroumeliotis/seomized/blob/main/seomized.sql> (accessed on 9 November 2023).
380. Organic Traffic and Why It Is Important Available online: <https://www.linkedin.com/pulse/organic-traffic-why-important-krohn-online-traffic-generation> (accessed on 9 November 2023).
381. Is Direct Traffic an Indicator of Brand Strength? - Portent Available online: <https://www.portent.com/blog/analytics/is-direct-traffic-an-indicator-of-brand-strength.htm> (accessed on 9 November 2023).
382. Egri, G.; Bayrak, C. The Role of Search Engine Optimization on Keeping the User on the Site. *Procedia Comput Sci* **2014**, 36, 335–342, doi:10.1016/J.PROCS.2014.09.102.
383. Organic Keywords: What Are They & How to Find Them Available online: <https://www.semrush.com/blog/organic-keywords/> (accessed on 9 November 2023).
384. Saura, J.R.; Palos-Sánchez, P.; Suárez, L.M.C. Understanding the Digital Marketing Environment with KPIs and Web Analytics. *Future Internet* 2017, Vol. 9, Page 76 **2017**, 9, 76, doi:10.3390/FI9040076.

385. Sakas, D.P.; Giannakopoulos, N.T. Harvesting Crowdsourcing Platforms' Traffic in Favour of Air Forwarders' Brand Name and Sustainability. *Sustainability* 2021, Vol. 13, Page 8222 **2021**, 13, 8222, doi:10.3390/SU13158222.
386. What Is Organic Search? Everything You Need to Know Available online: <https://ahrefs.com/blog/organic-search/> (accessed on 9 November 2023).
387. Ziakis, C.; Vlachopoulou, M.; Kyrkoudis, T.; Karagkiozidou, M. Important Factors for Improving Google Search Rank. *Future Internet* 2019, Vol. 11, Page 32 **2019**, 11, 32, doi:10.3390/FI11020032.
388. IATA - Current Airline Members Available online: <https://www.iata.org/en/about/members/airline-list/> (accessed on 9 November 2023).
389. GitHub - Kroumeliotis/Airlines-Seo Available online: <https://github.com/kroumeliotis/airlines-seo> (accessed on 9 November 2023).
390. Curl Available online: <https://curl.se/> (accessed on 9 November 2023).
391. Ubersuggest: Free Keyword Research Tool - Neil Patel Available online: <https://neilpatel.com/ubersuggest/> (accessed on 9 November 2023).
392. Downloading IBM SPSS Statistics 27 Available online: <https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-27> (accessed on 9 November 2023).
393. Airlines-Seo/Airlines.Xls at Main · Kroumeliotis/Airlines-Seo · GitHub Available online: <https://github.com/kroumeliotis/airlines-seo/blob/main/Airlines.xls> (accessed on 9 November 2023).
394. Ahrens, J.P. Visualization and Data Analysis at the Exascale. **2011**, doi:10.2172/1011053.
395. Marketing Research / William Zikmund, Steven D'Alessandro, Ben Lowe, Hume Winzar, Barry J. Babin - Catalogue | National Library of Australia Available online: <https://catalogue.nla.gov.au/catalog/7127119> (accessed on 9 November 2023).
396. Shrestha, S.K. Brand Loyalty of Baby Diaper Products. *Management Dynamics* **2018**, 21, 79–88, doi:10.3126/MD.V21I1.27049.
397. Tabachnick, B.G.; Fidell, L.S. Using Multivariate Statistics Title: Using Multivariate Statistics. **2019**.
398. Probability and Statistics for Scientists and Engineers / Rao V. Dukkipati. - Vanderbilt University Available online: [https://catalog.library.vanderbilt.edu/discovery/fulldisplay/alma991043268751403276/01VAN\\_INST:vanui](https://catalog.library.vanderbilt.edu/discovery/fulldisplay/alma991043268751403276/01VAN_INST:vanui) (accessed on 9 November 2023).
399. Asadoorian, M.O.; Kantarelis, Demetrius. Essentials of Inferential Statistics. **2005**, 286.
400. P-Value: What It Is, How to Calculate It, and Why It Matters Available online: <https://www.investopedia.com/terms/p/p-value.asp> (accessed on 9 November 2023).
401. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a World Beyond "p < 0.05." *American Statistician* **2019**, 73, 1–19, doi:10.1080/00031305.2019.1583913.
402. SPSS Statistics Tutorials and Statistical Guides | Laerd Statistics Available online: <https://statistics.laerd.com/> (accessed on 9 November 2023).
403. Statistical Methods for Comparing Regression Coefficients Between Models on JSTOR Available online: <https://www.jstor.org/stable/2782277> (accessed on 9 November 2023).
404. Independent Variable (Treatment Variable) Definition and Uses - Statistics How To Available online: <https://www.statisticshowto.com/independent-variable-definition/> (accessed on 28 November 2023).
405. Machine Learning and Knowledge Discovery in Databases. **2020**, 11907, doi:10.1007/978-3-030-46147-8.
406. Keim, D.A.; Hao, M.C.; Dayal, U.; Janetzko, H.; Bak, P. Generalized Scatter Plots. <http://dx.doi.org/10.1057/ivs.2009.34> **2009**, 9, 301–311, doi:10.1057/IVS.2009.34.

407. Sakas, D.P.; Giannakopoulos, N.T. Big Data Contribution in Desktop and Mobile Devices Comparison, Regarding Airlines' Digital Brand Name Effect. *Big Data and Cognitive Computing* 2021, Vol. 5, Page 48 **2021**, 5, 48, doi:10.3390/BDCC5040048.
408. Gray, S.A.; Gray, S.; Cox, L.J.; Henly-Shepard, S. Mental Modeler: A Fuzzy-Logic Cognitive Mapping Modeling Tool for Adaptive Environmental Management. *Proceedings of the Annual Hawaii International Conference on System Sciences* **2013**, 965–973, doi:10.1109/HICSS.2013.399.
409. Salmeron, J.L. Supporting Decision Makers with Fuzzy Cognitive Maps. *Research-Technology Management* **2009**, 52, 53–59, doi:10.1080/08956308.2009.11657569.
410. Brin, S.; Page, L. Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* **2012**, 56, 3825–3833, doi:10.1016/J.COMNET.2012.10.007.
411. Google Maps Platform | Google for Developers Available online: <https://developers.google.com/maps> (accessed on 9 November 2023).
412. GitHub - Rkonstadinos/Seo-Techniques-Hotels Available online: <https://github.com/rkonstadinos/seo-techniques-hotels> (accessed on 9 November 2023).
413. Seo-Techniques-Hotels/Hotels.Xlsx at Main · Rkonstadinos/Seo-Techniques-Hotels · GitHub Available online: <https://github.com/rkonstadinos/seo-techniques-hotels/blob/main/hotels.xlsx> (accessed on 9 November 2023).
414. Park, H.M. Univariate Analysis and Normality Test Using SAS, Stata, and SPSS. **2015**.
415. 1.3.5.11. Measures of Skewness and Kurtosis Available online: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm> (accessed on 9 November 2023).
416. Canatan, H. ScienceDirect World Conference on Technology, Innovation and Entrepreneurship Literature Search Consisting of the Areas of Six Sigma's Usage. *Procedia-Social and Behavioral Sciences* **2015**, 195, 695–704, doi:10.1016/j.sbspro.2015.06.160.
417. Zhao, C.; Lu, J.; Duan, F. Application and Research of SEO in the Development of Web2.0 Site. *2009 2nd International Symposium on Knowledge Acquisition and Modeling, KAM 2009* **2009**, 1, 236–238, doi:10.1109/KAM.2009.69.
418. Mittal, M.K.; Kirar, N.; Meena, J. Implementation of Search Engine Optimization: Through White Hat Techniques. *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018* **2018**, 674–678, doi:10.1109/ICACCCN.2018.8748337.
419. Introducing the Accelerated Mobile Pages Project, for a Faster, Open Mobile Web Available online: <https://blog.google/products/search/introducing-accelerated-mobile-pages/> (accessed on 9 November 2023).