



Τμήμα Ηλεκτρολόγων Μηχανικών
& Μηχανικών Υπολογιστών

**ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ**

ΠΜΣ : Τεχνολογίες και Υπηρεσίες Ευφυών
Συστημάτων Πληροφορικής και
Επικοινωνιών

Διπλωματική Εργασία

Τίτλος : Αξιολόγηση Μεθόδων
Επεξεργασίας Φυσικής Γλώσσας για την
Αυτόματη Κατανόηση Κειμένου

Θεοχάρης Ρασσιάς Α.Μ. 1020

Επιβλέπων Καθηγητής : Παναγιώτης Ζέρβας

Περίληψη

Ελληνικά

Στόχος αυτής της διπλωματικής είναι η χρήση αλγορίθμων αναγνώρισης οντοτήτων για την αναγνώριση δεξιοτήτων σε βιογραφικά. Έχουμε ένα σύνολο βιογραφικών σε ελεύθερο κείμενο στην Αγγλική γλώσσα, με το σώμα κειμένων να αποτελείται από 1530 έγγραφα. Ο σκοπός είναι να ανιχνεύσουμε δεξιότητες (skills) του υποψηφίου. Αυτή η διπλωματική εργασία έχει μεγάλη χρησιμότητα σε εταιρίες ώστε να διευκολύνουν τη διαδικασία της εύρεσης ενός κατάλληλου υποψηφίου για πρόσληψη σε μια θέση. Επεξηγηματικά, η Αξιολόγηση μεθόδων Φυσικής Γλώσσας (NLP), η χρησιμότητα της NLP στην αγορά εργασίας, η αναγνώριση ονομαστικών οντοτήτων (NER) και τα οφέλη τους στην αγορά εργασίας. Τα Μεγάλα γλωσσικά Μοντέλα (LLM), η εφαρμογή τους σε εργασίες NLP, και τα πλεονεκτήματα της χρήσης των LLM για NER. Το πρώτο σκέλος της διπλωματικής εισαγάγει σε θεωρητικό πλαίσιο τους παραπάνω όρους και τους αναλύει περαιτέρω. Στο δεύτερο σκέλος της διπλωματικής γίνεται εισαγωγή στο πρακτικό κομμάτι, παρουσιάζοντας τη μεθοδολογία και τη χρήση του GPT-NER. Μέσω των δεδομένων που εισαγάγονται, παρατηρούμε γίνεται χρήση γλωσσικών μοντέλων, όπως το ChatGPT μέσω του ChatGPT wrapper, για να έχουμε πρόσβαση προγραμματιστικά στο ChatGPT με γλώσσα Python. Ο σκοπός θα είναι η επεξεργασία του κειμένου και δημιουργία ετικετών, σύμφωνα με τη στρατηγική του GPT-NER. Γίνεται ανάλυση της διαδικασίας που ακολουθείται, του κώδικα που χρησιμοποιείται και παρουσιάζονται τα αποτελέσματα.

Λέξεις - κλειδιά : LLM, Machine Learning, NER, GPT

English

The aim of this thesis is to use entity recognition algorithms for skill recognition in biographies. We have a set of free-text resumes in English, with the text corpus consisting of 1530 documents. The purpose is to detect the skills of the candidate. This thesis is of great use to companies in order to facilitate the process of finding a suitable candidate for recruitment to a position. Explanation of Natural Language Processing (NLP) Methods Evaluation, the usefulness of NLP in the job market, Named Entity Recognition (NER) and its benefits in the job market. Large Language Models (LLMs), their application to NLP tasks, and the advantages of using LLMs for NER. The

first part of the thesis introduces the above terms in a theoretical framework and further analyzes them. The second part of the thesis introduces the practical part, presenting the methodology and the use of GPT-NER. Through the data introduced, we observe that it is made use of language models, such as ChatGPT through the ChatGPT wrapper, to access ChatGPT programmatically with Python language. The purpose will be to process the text and create tags according to the GPT-NER strategy. An analysis of the process followed, the code used and the results are presented.

Περιεχόμενα

1.	Εισαγωγή	6
1.1.	Ανάλυση δεδομένων αγοράς εργασίας.....	6
1.1.1.	Επισκόπηση της αγοράς εργασίας.....	6
1.1.2.	Σημασία της αναγνώρισης δεξιοτήτων.....	9
1.2.	Αξιοποίηση της NLP στην αγορά εργασίας.....	12
1.2.1.	Εφαρμογές του NLP στο HR.....	12
1.2.2.	Οφέλη του NLP για την αναγνώριση δεξιοτήτων	17
1.3.	Εργασίες NLP σχετικές με την ανάλυση της αγοράς εργασίας.....	21
1.3.1.	Ταξινόμηση κειμένου.....	21
1.3.2.	Αναγνώριση ονομαστικών οντοτήτων (NER)	25
1.3.3.	Εξαγωγή λέξεων-κλειδιών.....	28
1.4.	Λεπτομερής συζήτηση σχετικά με την NER στην ανάλυση της αγοράς εργασίας.....	32
1.4.1.	Τεχνικές NER για την εξαγωγή δεξιοτήτων	32
1.4.2.	Προκλήσεις στην NER για την ανάλυση βιογραφικού σημειώματος 38	
1.4.3.	Μελέτες περιπτώσεων και παραδείγματα	45
2.	Η κατάσταση της τέχνης.....	52
2.1.	Ανάλυση παραδοσιακών δεδομένων της αγοράς εργασίας.....	52
2.1.1	Ιστορικές προσεγγίσεις για την ανάλυση της αγοράς εργασίας.....	52
2.1.2	Περιορισμοί των παραδοσιακών μεθόδων	56
2.2	Εργασίες NLP στην ανάλυση της αγοράς εργασίας	60
2.2.1	Εξέλιξη της NLP στο HR	60
2.2.2	Τρέχουσες εφαρμογές NLP	65
2.3	Εργασία αναγνώρισης ονομαστικών οντοτήτων (NER).....	70
2.3.1	Βασικές αρχές της NER.....	70
2.3.2	NER στο πλαίσιο του HR	75
2.4	Μεγάλα γλωσσικά μοντέλα (LLM)	78
2.4.1	Εισαγωγή στα LLM.....	78

2.4.2 Εφαρμογές των LLM σε εργασίες NLP	80
2.5 NER με LLMs.....	84
2.5.1 Πλεονεκτήματα της χρήσης LLMs για NER	84
2.5.2 Μελέτες περιπτώσεων και εφαρμογές	87
2.5.3 Προκλήσεις και μελλοντικές κατευθύνσεις.....	90
3. Μεθοδολογία.....	95
3.1. Μοντέλο GPT-NER.....	95
4. Δεδομένα.....	98
5. Υλοποίηση και αποτελέσματα	100
5.1 ChatGPT.....	100
5.2 GPT4all	101
5.2.1. Εγκατάσταση	101
5.2.2. Κώδικας.....	101
5.3 ChatGPT API.....	103
5.3.1. Εγκατάσταση	103
5.3.2. Κώδικας.....	103
6. Αξιολόγηση	106
Βιβλιογραφία.....	109

1. Εισαγωγή

1.1. Ανάλυση δεδομένων αγοράς εργασίας

1.1.1. Επισκόπηση της αγοράς εργασίας

Τρέχουσες τάσεις

Η αγορά εργασίας υφίσταται σημαντικούς μετασχηματισμούς που οφείλονται σε διάφορους παράγοντες, συμπεριλαμβανομένων των τεχνολογικών εξελίξεων, των μετατοπίσεων των τρόπων εργασίας και της εξελισσόμενης φύσης των δομών απασχόλησης. Οι τάσεις αυτές επαναπροσδιορίζουν τη ζήτηση για δεξιότητες και αναδιαμορφώνουν το τοπίο του εργατικού δυναμικού (Milosevic et al., 2021).

Μια από τις πιο εμφανείς τάσεις είναι η αυξανόμενη ζήτηση για δεξιότητες που σχετίζονται με την τεχνολογία. Καθώς οι βιομηχανίες συνεχίζουν να ψηφιοποιούνται, η ανάγκη για τεχνογνωσία σε τομείς όπως η ανάπτυξη λογισμικού, η ανάλυση δεδομένων, η κυβερνοασφάλεια και η τεχνητή νοημοσύνη έχει αυξηθεί κατακόρυφα. Οι εταιρείες βασίζονται ολοένα και περισσότερο σε διαδικασίες λήψης αποφάσεων με βάση τα δεδομένα, γεγονός που απαιτεί ένα εργατικό δυναμικό ικανό στο χειρισμό μεγάλων συνόλων δεδομένων και στην εξαγωγή αξιοποιήσιμων πληροφοριών. Ο πολλαπλασιασμός των ψηφιακών τεχνολογιών σε όλους τους τομείς έχει καταστήσει τον τεχνολογικό αλφαριθμητισμό όχι απλώς πλεονέκτημα αλλά αναγκαιότητα. Η μετατόπιση αυτή είναι ιδιαίτερα εμφανής σε τομείς όπως η χρηματοδότηση, η υγειονομική περίθαλψη και η μεταποίηση, όπου η τεχνολογική ολοκλήρωση είναι απαραίτητη για τη διατήρηση της ανταγωνιστικότητας και της αποτελεσματικότητας (Michaels et al., 2014).

Η αύξηση της εξ αποστάσεως εργασίας επηρέασε περαιτέρω τις απαιτήσεις δεξιοτήτων. Η πανδημία COVID-19 επιτάχυνε την υιοθέτηση της εξ αποστάσεως εργασίας, μετατρέποντάς την από σχέδιο έκτακτης ανάγκης σε βασικό τρόπο εργασίας. Η μετατόπιση αυτή ανέδειξε τη σημασία των ψηφιακών εργαλείων επικοινωνίας και συνεργασίας, καθώς και των μέτρων κυβερνοασφάλειας για την προστασία των απομακρυσμένων εργασιών. Οι εργαζόμενοι πρέπει πλέον να είναι έμπειροι στη χρήση διαφόρων διαδικτυακών πλατφορμών για συσκέψεις, διαχείριση έργων και καθημερινή επικοινωνία (Apatsidis et al., 2021). Επιπλέον, η ικανότητα ανεξάρτητης εργασίας και αποτελεσματικής διαχείρισης του χρόνου τους έχει καταστεί ζωτικής σημασίας. Η απομακρυσμένη εργασία έχει επίσης διευρύνει τη δεξαμενή ταλέντων για τους εργοδότες, επιτρέποντάς τους να προσλαμβάνουν από μια παγκόσμια αγορά. Ωστόσο, αυτό έχει επίσης εντείνει τον ανταγωνισμό μεταξύ των αιτούντων εργασία, οι οποίοι πρέπει πλέον να

αποδείξουν όχι μόνο τις τεχνικές τους δεξιότητες αλλά και την ικανότητά τους να ευδοκιμούν σε ένα περιβάλλον εργασίας εξ αποστάσεως (Avdeeva et al., 2021).

Ειδικές απαιτήσεις της βιομηχανίας

Το τοπίο της αγοράς εργασίας διαμορφώνεται περίπλοκα από τις μοναδικές απαιτήσεις των διαφόρων κλάδων, καθένας από τους οποίους καθοδηγείται από τις ξεχωριστές λειτουργικές του ανάγκες και τις τεχνολογικές εξελίξεις. Μεταξύ αυτών, οι τομείς της υγείας, της χρηματοδότησης και της μεταποίησης παρουσιάζουν σε περίοπτη θέση τις εξελισσόμενες απαιτήσεις δεξιοτήτων που υπαγορεύουν τις τρέχουσες και μελλοντικές τάσεις απασχόλησης (Shmatko & Volkova, 2020).

Στον κλάδο της υγειονομικής περίθαλψης, υπάρχει αυξανόμενη ανάγκη τόσο για παραδοσιακούς επαγγελματίες του ιατρικού κλάδου όσο και για άτομα που είναι έμπειρα στην τεχνολογία πληροφοριών (ΤΠ) της υγειονομικής περίθαλψης. Η γήρανση του πληθυσμού και οι συνεχιζόμενες παγκόσμιες προκλήσεις στον τομέα της υγείας έχουν ενισχύσει τη ζήτηση για γιατρούς, νοσηλευτές και επαγγελματίες υγείας. Ωστόσο, η ζήτηση αυτή επεκτείνεται πέρα από τον κλινικό τομέα, υπογραμμίζοντας την ανάγκη για τεχνογνωσία στην πληροφορική της υγειονομικής περίθαλψης. Η ενσωμάτωση των ηλεκτρονικών φακέλων υγείας (ΗΦΥ), της τηλεϊατρικής και της ανάλυσης δεδομένων στην παροχή υγειονομικής περίθαλψης έχει φέρει επανάσταση στη φροντίδα των ασθενών. Οι επαγγελματίες με εξειδίκευση στην πληροφορική της υγείας, τη διαχείριση δεδομένων και την κυβερνοασφάλεια είναι απαραίτητοι για να διασφαλίσουν την αποτελεσματική εφαρμογή και την ασφαλή διαχείριση αυτών των τεχνολογιών. Η ικανότητα ανάλυσης των δεδομένων υγείας για τη βελτίωση των αποτελεσμάτων των ασθενών και της επιχειρησιακής αποδοτικότητας καθίσταται ολοένα και πιο ζωτικής σημασίας, γεγονός που απαιτεί ένα εργατικό δυναμικό ικανό τόσο στις ιατρικές γνώσεις όσο και στις δεξιότητες πληροφορικής (Jurczuk & Florea, 2022).

Ο χρηματοπιστωτικός κλάδος βιώνει μια παρόμοια εξέλιξη, η οποία οφείλεται στη ραγδαία εξέλιξη της χρηματοοικονομικής τεχνολογίας (fintech). Οι παραδοσιακές δεξιότητες που απαιτούνται στα χρηματοοικονομικά, όπως η λογιστική και η χρηματοοικονομική ανάλυση, συμπληρώνονται τώρα από την ανάγκη για επάρκεια στις εφαρμογές fintech. Ο ψηφιακός μετασχηματισμός στα χρηματοοικονομικά έχει οδηγήσει σε αύξηση της ζήτησης για δεξιότητες που σχετίζονται με την τεχνολογία blockchain, την τεχνητή νοημοσύνη και τη μηχανική μάθηση (Shmatko & Volkova, 2020). Οι τεχνολογίες αυτές είναι καθοριστικής σημασίας για την ανάπτυξη καινοτόμων χρηματοοικονομικών προϊόντων, την ενίσχυση των μέτρων ασφαλείας και τη βελτιστοποίηση των διαδικασιών συναλλαγών. Επιπλέον, η διαχείριση κινδύνων και η συμμόρφωση

έχουν καταστεί κρίσιμοι τομείς εστίασης λόγω του αυστηρού ρυθμιστικού περιβάλλοντος και της αυξανόμενης πολυπλοκότητας των χρηματοπιστωτικών προϊόντων. Οι επαγγελματίες που μπορούν να περιηγηθούν στο ρυθμιστικό τοπίο, να διεξάγουν ενδελεχείς αξιολογήσεις κινδύνου και να διασφαλίζουν τη συμμόρφωση με τους χρηματοοικονομικούς νόμους και κανονισμούς είναι ιδιαίτερα περιζήτητοι. Η ικανότητα αξιοποίησης της τεχνολογίας για την ανάλυση και τον μετριασμό των κινδύνων αποτελεί βασικό πλεονέκτημα στον σύγχρονο χρηματοπιστωτικό τομέα (Li et al., 2021).

Περιφερειακές παραλλαγές

Η αγορά εργασίας δεν είναι μια μονολιθική οντότητα- διαφέρει σημαντικά σε διάφορες περιοχές και χώρες λόγω των διαφορετικών οικονομικών δομών, των πολιτισμικών πλαισίων και των αναπτυξιακών σταδίων. Αυτές οι περιφερειακές διαφοροποιήσεις επηρεάζουν σε μεγάλο βαθμό τη ζήτηση για δεξιότητες, διαμορφώνοντας ξεχωριστά τοπία απασχόλησης που αντικατοπτρίζουν τις τοπικές ανάγκες και ευκαιρίες (Bocharova, 2020).

Η σύγκριση των απαιτήσεων σε δεξιότητες μεταξύ διαφορετικών περιοχών ή χωρών αποκαλύπτει έντονες αντιθέσεις που έχουν τις ρίζες τους στην οικονομική εξειδίκευση και στα επίπεδα ανάπτυξης. Στις τεχνολογικά προηγμένες οικονομίες, όπως οι Ηνωμένες Πολιτείες, η Γερμανία και η Ιαπωνία, υπάρχει υψηλή ζήτηση για προηγμένες τεχνολογικές δεξιότητες, όπως η μηχανική λογισμικού, η επιστήμη των δεδομένων και η τεχνητή νοημοσύνη. Αυτά τα έθνη επενδύουν σε μεγάλο βαθμό σε τομείς με γνώμονα την καινοτομία, απαιτώντας ένα εργατικό δυναμικό ικανό σε τεχνολογίες αιχμής για να διατηρήσουν το ανταγωνιστικό τους πλεονέκτημα. Αντίθετα, οι αναδυόμενες οικονομίες, όπως η Ινδία, η Βραζιλία και η Ινδονησία, παρουσιάζουν σημαντική ζήτηση για δεξιότητες στον τομέα της μεταποίησης, της γεωργίας και των βασικών υπηρεσιών. Ενώ οι χώρες αυτές αντιμετωπίζουν επίσης μια αυξανόμενη ανάγκη για δεξιότητες που σχετίζονται με την τεχνολογία, ο ρυθμός της αλλαγής και η κλίμακα της ζήτησης διαφέρουν σημαντικά σε σύγκριση με τις πιο ανεπτυγμένες περιοχές (Young, 2013).

Τα εκπαιδευτικά συστήματα και τα προγράμματα επαγγελματικής κατάρτισης στις διάφορες περιοχές επηρεάζουν επίσης τη ζήτηση δεξιοτήτων. Για παράδειγμα, χώρες με ισχυρά συστήματα μαθητείας, όπως η Γερμανία και η Ελβετία, τείνουν να έχουν σταθερή προσφορά εξειδικευμένων επαγγελματιών, καλύπτοντας τις ανάγκες των βιομηχανικών και μεταποιητικών τους τομέων. Αντίθετα, περιοχές με λιγότερη έμφαση στην επαγγελματική κατάρτιση μπορεί να αντιμετωπίζουν ελλείψεις σε συγκεκριμένους τομείς δεξιοτήτων, επηρεάζοντας την οικονομική τους ανάπτυξη και τη σταθερότητα της αγοράς εργασίας. Επιπλέον, το πολιτικό περιβάλλον και η κυβερνητική υποστήριξη συγκεκριμένων βιομηχανιών διαδραματίζουν καθοριστικό ρόλο στη διαμόρφωση

των περιφερειακών απαιτήσεων σε δεξιότητες. Περιφέρειες που προωθούν ενεργά τομείς όπως οι ανανεώσιμες πηγές ενέργειας, η βιοτεχνολογία ή η τεχνολογία της πληροφορίας μέσω επιδοτήσεων και κινήτρων συχνά βλέπουν αντίστοιχη αύξηση της ζήτησης για εξειδικευμένες δεξιότητες στους τομείς αυτούς (Bocharova, 2020).

1.1.2. Σημασία της αναγνώρισης δεξιοτήτων

Προοπτική του εργοδότη

Στη σύγχρονη αγορά εργασίας, η αναγνώριση των δεξιοτήτων έχει καταστεί κρίσιμος παράγοντας τόσο για τους εργοδότες όσο και για τους αναζητούντες εργασία. Για τους εργοδότες, η ικανότητα αποτελεσματικού εντοπισμού και αντιστοίχισης των κατάλληλων υποψηφίων με τις κενές θέσεις εργασίας είναι υψίστης σημασίας. Η αυξανόμενη πολυπλοκότητα και η εξειδίκευση πολλών ρόλων σημαίνει ότι δεν αρκεί πλέον η απλή κατοχή ενός πτυχίου ή ενός συγκεκριμένου αριθμού εμπειρίας- οι συγκεκριμένες δεξιότητες αποτελούν πλέον τους βασικούς δείκτες της καταλληλότητας ενός υποψηφίου για μια θέση. Κατά συνέπεια, οι διαδικασίες αναγνώρισης δεξιοτήτων διαδραματίζουν καθοριστικό ρόλο στον εξορθολογισμό των προσλήψεων, στη μείωση του κόστους και στη βελτίωση της συνολικής οργανωτικής αποτελεσματικότητας (Petrican et al., 2017).

Από τη σκοπιά του εργοδότη, ένα από τα κύρια οφέλη της αποτελεσματικής αναγνώρισης δεξιοτήτων είναι η αυξημένη αποτελεσματικότητα στην αντιστοίχιση των κενών θέσεων εργασίας με τους σωστούς υποψηφίους. Οι παραδοσιακές διαδικασίες πρόσληψης συχνά περιλαμβάνουν σημαντική χειρωνακτική προσπάθεια, όπου οι υπεύθυνοι προσλήψεων κοσκινίζουν πολυάριθμα βιογραφικά για να εντοπίσουν πιθανούς υποψηφίους. Αυτή η προσέγγιση δεν είναι μόνο χρονοβόρα αλλά και επιρρεπής σε σφάλματα και προκαταλήψεις (Erdem, 2023). Τα αυτοματοποιημένα συστήματα αναγνώρισης δεξιοτήτων, που υποστηρίζονται από προηγμένες τεχνικές επεξεργασίας φυσικής γλώσσας (NLP), έχουν φέρει επανάσταση σε αυτή τη διαδικασία. Αυτά τα συστήματα μπορούν να αναλύουν γρήγορα βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας για να εντοπίζουν τις βασικές δεξιότητες και προσόντα, διασφαλίζοντας ότι εξετάζονται μόνο οι πιο κατάλληλοι υποψήφιοι. Αυτή η αυτοματοποιημένη διαδικασία αντιστοίχισης μειώνει σημαντικά τον φόρτο εργασίας των ανθρώπινων στελεχών προσλήψεων, επιτρέποντάς τους να επικεντρωθούν σε πιο στρατηγικές πτυχές της απόκτησης ταλέντων (Chala et al., 2017).

Επιπλέον, η αύξηση της αποδοτικότητας από την αυτοματοποιημένη αναγνώριση δεξιοτήτων μεταφράζεται άμεσα σε μείωση του χρόνου πρόσληψης. Ο παραδοσιακός κύκλος προσλήψεων, από την ανάρτηση θέσεων εργασίας έως την επιλογή υποψηφίων, μπορεί να είναι χρονοβόρος και συχνά διαρκεί αρκετές εβδομάδες ή ακόμη και μήνες. Κατά τη διάρκεια αυτής της περιόδου, οι επιχειρήσεις μπορεί να υποφέρουν από μειωμένη παραγωγικότητα και την πίεση των ακάλυπτων θέσεων (Jayasinghe et al., 2022). Τα αυτοματοποιημένα συστήματα επιταχύνουν την αρχική φάση ελέγχου, περιορίζοντας γρήγορα τη δεξαμενή των υποψηφίων σε εκείνους που πληρούν τα καθορισμένα κριτήρια. Συντομεύοντας τον χρόνο που απαιτείται για τον εντοπισμό των κατάλληλων υποψηφίων, οι επιχειρήσεις μπορούν να καλύπτουν τις κενές θέσεις πιο γρήγορα, ελαχιστοποιώντας τη διακοπή της λειτουργίας τους. Αυτή η ταχεία διαδικασία πρόσληψης είναι ιδιαίτερα επωφελής σε κλάδους με υψηλά ποσοστά κύκλου εργασιών ή σε κλάδους που παρουσιάζουν ταχεία ανάπτυξη, όπου η ικανότητα ταχείας πρόσληψης είναι απαραίτητη για τη διατήρηση του ανταγωνιστικού πλεονεκτήματος (Erdem, 2023).

Προοπτική του αναζητούντος εργασία

Από τη σκοπιά των ατόμων που αναζητούν εργασία, η σημασία της αναγνώρισης δεξιοτήτων στην αγορά εργασίας δεν μπορεί να υπερεκτιμηθεί. Σε ένα ανταγωνιστικό τοπίο εργασίας, όπου οι υποψήφιοι διεκδικούν την προσοχή των εργοδοτών και των υπεύθυνων προσλήψεων, η ικανότητα σαφούς διατύπωσης και απόδειξης των δεξιοτήτων ενός ατόμου μπορεί να ενισχύσει σημαντικά την προβολή και να βελτιώσει τα αποτελέσματα της εύρεσης εργασίας. Τα συστήματα αναγνώρισης δεξιοτήτων, τα οποία χρησιμοποιούν προηγμένους αλγορίθμους για την ανάλυση και την ανάδειξη των σχετικών ικανοτήτων, παρέχουν στους αναζητούντες εργασία ένα κρίσιμο πλεονέκτημα, διασφαλίζοντας ότι οι ικανότητές τους αντιπροσωπεύονται με ακρίβεια και ευθυγραμμίζονται με τις απαιτήσεις των πιθανών εργοδοτών (Chala et al., 2017). Ένα από τα κύρια οφέλη της αναγνώρισης δεξιοτήτων για τους αναζητούντες εργασία είναι η ενισχυμένη προβολή των σχετικών δεξιοτήτων τους. Σε μια παραδοσιακή διαδικασία υποβολής αιτήσεων εργασίας, τα βιογραφικά σημειώματα και οι συνοδευτικές επιστολές συχνά ξεετάζονται χειροκίνητα από τους υπεύθυνους προσλήψεων, οι οποίοι ενδέχεται να παραβλέψουν βασικά προσόντα λόγω του μεγάλου όγκου των αιτήσεων. Τα αυτοματοποιημένα συστήματα αναγνώρισης δεξιοτήτων αμβλύνουν αυτό το πρόβλημα με τον συστηματικό εντοπισμό και την επισήμανση των σχετικών δεξιοτήτων σε ένα βιογραφικό σημείωμα, διασφαλίζοντας ότι τα προσόντα αυτά εμφανίζονται σε περίοπτη θέση. Αυτή η αυξημένη ορατότητα σημαίνει ότι οι αιτούντες εργασία είναι πιο πιθανό να τραβήξουν την προσοχή των υπεύθυνων προσλήψεων και

των διευθυντών προσλήψεων, αυξάνοντας έτσι τις πιθανότητές τους να εξεταστούν για τις σχετικές θέσεις. Εξασφαλίζοντας ότι οι πιο κρίσιμες πτυχές των δεξιοτήτων ενός υποψηφίου δεν είναι θαμμένες μέσα σε πυκνό κείμενο, τα συστήματα αυτά βοηθούν τους αιτούντες εργασία να παρουσιάζονται πιο αποτελεσματικά και να ευθυγραμμίζουν τα προσόντα τους με τις απαιτήσεις της θέσης εργασίας (Yang et al., 2022).

Επιπλέον, η βελτιωμένη ακρίβεια στην ανάδειξη των βασικών δεξιοτήτων μεταφράζεται άμεσα σε καλύτερες αντιστοιχίες θέσεων εργασίας. Όταν οι δεξιότητες των υποψηφίων αναγνωρίζονται με ακρίβεια και αντιστοιχίζονται με τις περιγραφές θέσεων εργασίας, αυξάνεται η πιθανότητα να εξεταστούν για θέσεις που πραγματικά ταιριάζουν στα προσόντα και τις επαγγελματικές τους φιλοδοξίες (Chala et al., 2017). Αυτή η ακριβής ευθυγράμμιση δεν ωφελεί μόνο τους εργοδότες, αλλά διασφαλίζει επίσης ότι οι υποψήφιοι για εργασία δεν χάνουν χρόνο κάνοντας αίτηση για ρόλους που δεν ταιριάζουν στις δεξιότητές τους. Κατά συνέπεια, οι αναζητούντες εργασία μπορούν να επικεντρώσουν τις προσπάθειές τους σε ευκαιρίες που είναι πιο πιθανό να οδηγήσουν σε επιτυχή απασχόληση, ενισχύοντας έτσι την αποτελεσματικότητα της αναζήτησης εργασίας και τη συνολική ικανοποίησή τους. Αυτή η ευθυγράμμιση μεταξύ των δεξιοτήτων και των απαιτήσεων της εργασίας είναι ιδιαίτερα κρίσιμη σε εξειδικευμένους τομείς όπου οι συγκεκριμένες τεχνικές ικανότητες είναι υψίστης σημασίας (Drewery et al., 2022).

Οικονομικός αντίκτυπος

Ο οικονομικός αντίκτυπος της αναγνώρισης δεξιοτήτων είναι βαθύς και πολύπλευρος, επηρεάζοντας τόσο τα ατομικά αποτελέσματα της σταδιοδρομίας όσο και τις ευρύτερες οικονομικές τάσεις. Η ευθυγράμμιση των δεξιοτήτων του εργατικού δυναμικού με τις ανάγκες της αγοράς είναι απαραίτητη για την οικονομική αποτελεσματικότητα και παραγωγικότητα. Τα συστήματα αναγνώρισης δεξιοτήτων διαδραματίζουν κρίσιμο ρόλο στη διασφάλιση ότι οι δεξιότητες του εργατικού δυναμικού ανταποκρίνονται στις εξελισσόμενες απαιτήσεις των βιομηχανιών, συμβάλλοντας έτσι σημαντικά στην οικονομική σταθερότητα και ανάπτυξη (Dikshit & Jain, 2017).

Ένα από τα κύρια οικονομικά οφέλη της αποτελεσματικής αναγνώρισης δεξιοτήτων είναι η καλύτερη ευθυγράμμιση των δεξιοτήτων του εργατικού δυναμικού με τις ανάγκες της αγοράς. Καθώς οι κλάδοι εξελίσσονται και εμφανίζονται νέες τεχνολογίες, η ζήτηση για συγκεκριμένες δεξιότητες αλλάζει γρήγορα (Morris et al., 2019). Τα παραδοσιακά προγράμματα εκπαίδευσης και κατάρτισης συχνά δυσκολεύονται να συμβαδίσουν με αυτές τις αλλαγές, με αποτέλεσμα να υπάρχει αναντιστοιχία μεταξύ των δεξιοτήτων που διαθέτει το εργατικό δυναμικό και εκείνων που απαιτούνται από τους εργοδότες. Τα

συστήματα αναγνώρισης δεξιοτήτων αντιμετωπίζουν αυτό το ζήτημα με τον ακριβή προσδιορισμό και την κατηγοριοποίηση των δεξιοτήτων που υπάρχουν στο εργατικό δυναμικό. Παρέχοντας μια σαφή εικόνα των διαθέσιμων δεξιοτήτων, τα συστήματα αυτά επιτρέπουν στους εργοδότες να λαμβάνουν πιο τεκμηριωμένες αποφάσεις πρόσληψης και να εντοπίζουν πιο αποτελεσματικά τους κατάλληλους υποψηφίους. Η ευθυγράμμιση αυτή μειώνει τον χρόνο και τους πόρους που δαπανώνται για την εκπαίδευση νέων προσλήψεων, ενισχύοντας έτσι την παραγωγικότητα του οργανισμού και την οικονομική αποδοτικότητα (Puerta et al., 2018).

Επιπλέον, η αναγνώριση δεξιοτήτων διευκολύνει τη δυναμική ανακατανομή της εργασίας εντός της οικονομίας. Καθώς ορισμένοι κλάδοι επεκτείνονται και άλλοι συρρικνώνονται, η ικανότητα ταχείας αναγνώρισης μεταβιβάσιμων δεξιοτήτων αποκτά ολοένα και μεγαλύτερη σημασία. Οι εργαζόμενοι των οποίων οι κλάδοι φθίνουν μπορούν να επανεκπαιδευτούν και να μεταταχθούν σε αναπτυσσόμενους τομείς, ελαχιστοποιώντας την ανεργία και την υποαπασχόληση. Για παράδειγμα, η μετάβαση από την παραδοσιακή μεταποίηση στις προηγμένες τεχνολογίες μεταποίησης απαιτεί από τους εργαζόμενους να αποκτήσουν νέες δεξιότητες στον τομέα της αυτοματοποίησης και της ρομποτικής (Ejiwale, 2019). Τα συστήματα αναγνώρισης δεξιοτήτων μπορούν να βοηθήσουν στον εντοπισμό εργαζομένων με θεμελιώδεις δεξιότητες στη μεταποίηση, οι οποίοι μπορούν να επανεκπαιδευτούν για αυτές τις νέες τεχνολογίες. Αυτή η ρευστή μετακίνηση του εργατικού δυναμικού διασφαλίζει ότι η οικονομία μπορεί να προσαρμοστεί στις διαρθρωτικές αλλαγές και να συνεχίσει να αναπτύσσεται παρά τις τομεακές μετατοπίσεις (Morris et al., 2019).

1.2. Αξιοποίηση της NLP στην αγορά εργασίας

1.2.1. Εφαρμογές του NLP στο HR

Διαλογή βιογραφικού σημειώματος

Η ενσωμάτωση της Επεξεργασίας Φυσικής Γλώσσας (NLP) στους ανθρώπινους πόρους (HR) έχει φέρει επανάσταση στις παραδοσιακές διαδικασίες πρόσληψης, ιδίως μέσω της αυτοματοποίησης του ελέγχου βιογραφικών σημειωμάτων. Καθώς οι οργανισμοί αντιμετωπίζουν αυξανόμενο όγκο αιτήσεων εργασίας, η ανάγκη για αποτελεσματικές και ακριβείς μεθόδους αξιολόγησης των υποψηφίων έχει καταστεί υψίστης σημασίας. Η αυτοματοποιημένη ανάλυση των βιογραφικών σημειωμάτων για την εξαγωγή σχετικών πληροφοριών είναι μία από τις πιο επιδραστικές εφαρμογές της NLP στο HR, παρέχοντας σημαντικά

οφέλη όσον αφορά την αποτελεσματικότητα, την ακρίβεια και την επεκτασιμότητα (Ali et al., 2022).

Ο αυτοματοποιημένος έλεγχος βιογραφικών σημειωμάτων με τη χρήση NLP περιλαμβάνει τη χρήση εξελιγμένων αλγορίθμων για την ανάλυση βιογραφικών σημειωμάτων και την εξαγωγή σχετικών πληροφοριών, όπως η εργασιακή εμπειρία, το εκπαιδευτικό υπόβαθρο, οι δεξιότητες και άλλα προσόντα. Αυτή η τεχνολογία μιμείται την ανθρώπινη ικανότητα κατανόησης και ερμηνείας κειμένου, αλλά το κάνει με πολύ ταχύτερο ρυθμό και με μεγαλύτερη συνέπεια. Με την ανάλυση των βιογραφικών σημειωμάτων, τα συστήματα NLP μπορούν να εντοπίζουν τα βασικά χαρακτηριστικά και να τα κατηγοριοποιούν, επιτρέποντας στους υπεύθυνους προσλήψεων να εντοπίζουν γρήγορα τους καταλληλότερους υποψηφίους για μια συγκεκριμένη θέση (Korraparu, 2010).

Ένα από τα κύρια εργαλεία που χρησιμοποιούνται για τον αυτοματοποιημένο έλεγχο βιογραφικών σημειωμάτων είναι το σύστημα παρακολούθησης αιτούντων (ATS). Εταιρείες όπως το LinkedIn και το Indeed έχουν υιοθετήσει ευρέως λύσεις ATS για τον εξορθολογισμό των διαδικασιών πρόσληψής τους. Αυτά τα συστήματα χρησιμοποιούν NLP για να κοσκινίζουν μεγάλους όγκους βιογραφικών σημειωμάτων, να εξάγουν σχετικές πληροφορίες και να κατατάσσουν τους υποψηφίους με βάση το πόσο καλά τα προσόντα τους ταιριάζουν με τις απαιτήσεις της θέσης εργασίας. Αυτή η αυτοματοποιημένη προσέγγιση μειώνει σημαντικά το χειροκίνητο φόρτο εργασίας για τους επαγγελματίες του ανθρώπινου δυναμικού, επιτρέποντάς τους να επικεντρωθούν σε πιο στρατηγικές πτυχές της πρόσληψης, όπως οι συνεντεύξεις και η δέσμευση των υποψηφίων (Lalitha et al., 2023).

Η χρήση της ATS με τη βοήθεια του NLP προσφέρει πολλά πλεονεκτήματα. Πρώτον, ενισχύει την αποτελεσματικότητα της διαδικασίας ελέγχου βιογραφικών σημειωμάτων. Οι παραδοσιακές χειροκίνητες μέθοδοι διαλογής είναι χρονοβόρες και επιρρεπείς σε ανθρώπινα λάθη. Οι υπεύθυνοι προσλήψεων μπορούν να εξετάσουν μόνο έναν περιορισμένο αριθμό βιογραφικών σημειωμάτων σε ένα δεδομένο χρονικό διάστημα, γεγονός που μπορεί να έχει ως αποτέλεσμα να παραβλέπονται υποψήφιοι με προσόντα. Αντίθετα, ένα ATS μπορεί να επεξεργαστεί χιλιάδες βιογραφικά σε ένα κλάσμα του χρόνου, διασφαλίζοντας ότι κανένας δυνητικά κατάλληλος υποψήφιος δεν θα χαθεί. Αυτή η ταχύτητα είναι ιδιαίτερα κρίσιμη για μεγάλους οργανισμούς που λαμβάνουν μεγάλο όγκο αιτήσεων για κάθε θέση εργασίας (Nimbekar et al., 2019).

Δεύτερον, ο έλεγχος βιογραφικών σημειωμάτων με βάση τη NLP βελτιώνει την ακρίβεια της επιλογής υποψηφίων. Οι άνθρωποι υπεύθυνοι προσλήψεων, παρά την εμπειρογνομοσύνη τους, ενδέχεται να εισάγουν ακούσια προκαταλήψεις ή ασυνέπειες στη διαδικασία διαλογής. Παράγοντες όπως η κόπωση, οι

ασυνείδητες προκαταλήψεις και η υποκειμενική κρίση μπορούν να επηρεάσουν τη λήψη αποφάσεων (Lalitha et al., 2023). Τα αυτοματοποιημένα συστήματα, από την άλλη πλευρά, εφαρμόζουν συνεπή κριτήρια σε όλα τα βιογραφικά σημειώματα, εξαλείφοντας αυτές τις προκαταλήψεις και εξασφαλίζοντας μια πιο δίκαιη διαδικασία αξιολόγησης. Αυτή η αντικειμενικότητα βοηθά τους οργανισμούς να δημιουργήσουν ένα πιο ποικιλόμορφο και χωρίς αποκλεισμούς εργατικό δυναμικό, εστιάζοντας αποκλειστικά στα προσόντα και τις δεξιότητες των υποψηφίων (Korraparu, 2010).

Επιπλέον, η επεκτασιμότητα του ελέγχου βιογραφικών σημειωμάτων με βάση τη NLP είναι ένα άλλο σημαντικό πλεονέκτημα. Καθώς οι εταιρείες μεγαλώνουν και επεκτείνονται, οι ανάγκες τους για προσλήψεις αυξάνονται αντίστοιχα. Η κλιμάκωση των χειροκίνητων διαδικασιών διαλογής είναι συχνά μη πρακτική και δαπανηρή. Ένα ATS, ωστόσο, μπορεί να διαχειριστεί έναν αυξανόμενο όγκο αιτήσεων χωρίς πρόσθετη δαπάνη πόρων. Αυτή η επεκτασιμότητα διασφαλίζει ότι οι διαδικασίες πρόσληψης παραμένουν αποδοτικές και αποτελεσματικές, ανεξάρτητα από το μέγεθος του οργανισμού ή τον αριθμό των θέσεων που καλύπτονται (Nimbekar et al., 2019).

Αντιστοίχιση θέσεων εργασίας

Η χρήση της επεξεργασίας φυσικής γλώσσας (NLP) στο ανθρώπινο δυναμικό (HR) εκτείνεται πέρα από τον έλεγχο βιογραφικών σημειωμάτων και περιλαμβάνει αλγόριθμους που αντιστοιχούν τα προφίλ των υποψηφίων με τις περιγραφές των θέσεων εργασίας. Αυτή η εφαρμογή του NLP βελτιώνει σημαντικά τη διαδικασία πρόσληψης, διασφαλίζοντας ότι οι καλύτεροι υποψήφιοι αντιστοιχίζονται με τις κατάλληλες ευκαιρίες απασχόλησης, βελτιστοποιώντας έτσι τις εμπειρίες τόσο του εργοδότη όσο και του αιτούντος εργασία (Pendyala et al., 2022).

Η αντιστοίχιση των προφίλ των υποψηφίων με τις περιγραφές των θέσεων εργασίας περιλαμβάνει εξελιγμένους αλγόριθμους NLP που αναλύουν το κείμενο των αγγελιών και των βιογραφικών σημειωμάτων των υποψηφίων για να εντοπίσουν όρους και φράσεις-κλειδιά. Αυτοί οι αλγόριθμοι εξετάζουν διάφορους παράγοντες, όπως δεξιότητες, προσόντα, τίτλους εργασίας και επίπεδα εμπειρίας, για να καθορίσουν τη συμβατότητα μεταξύ ενός υποψηφίου και μιας θέσης εργασίας. Η διαδικασία ξεκινά με την ανάλυση τόσο των περιγραφών θέσεων εργασίας όσο και των προφίλ των υποψηφίων για την εξαγωγή σχετικών πληροφοριών, οι οποίες στη συνέχεια συγκρίνονται με τη χρήση προηγμένων τεχνικών αντιστοίχισης. Αυτή η αυτοματοποιημένη προσέγγιση εξασφαλίζει υψηλό βαθμό ακρίβειας και αποτελεσματικότητας, μειώνοντας σημαντικά τον χρόνο και την προσπάθεια που απαιτείται για τη χειροκίνητη αντιστοίχιση (Deng et al., 2018).

Πλατφόρμες όπως η Glassdoor και η ZipRecruiter αποτελούν παράδειγμα της επιτυχημένης εφαρμογής αλγορίθμων NLP στην αντιστοίχιση προφίλ υποψηφίων με περιγραφές θέσεων εργασίας. Η Glassdoor, γνωστή για τις ολοκληρωμένες λειτουργίες αναζήτησης θέσεων εργασίας και αξιολόγησης εταιρειών, χρησιμοποιεί NLP για να παρέχει εξατομικευμένες συστάσεις θέσεων εργασίας στους χρήστες. Αναλύοντας το περιεχόμενο των δημοσιεύσεων θέσεων εργασίας και των προφίλ των χρηστών, οι αλγόριθμοι της Glassdoor μπορούν να προτείνουν ευκαιρίες απασχόλησης που ταιριάζουν στενά με τις δεξιότητες και τις επαγγελματικές φιλοδοξίες ενός υποψηφίου. Αυτή η εξατομικευμένη διαδικασία αντιστοίχισης όχι μόνο αυξάνει την πιθανότητα εύρεσης κατάλληλων ευκαιριών απασχόλησης, αλλά και ενισχύει την ικανοποίηση των χρηστών παρέχοντας σχετικές και έγκαιρες προτάσεις (Martinez-Gil et al., 2016).

Ομοίως, το ZipRecruiter χρησιμοποιεί προηγμένους αλγόριθμους NLP για να διευκολύνει την αποτελεσματική αντιστοίχιση θέσεων εργασίας. Η τεχνολογία αντιστοίχισης της πλατφόρμας με βάση την τεχνητή νοημοσύνη αναλύει καθημερινά εκατομμύρια δημοσιεύσεις θέσεων εργασίας και προφίλ υποψηφίων. Αξιοποιώντας τη NLP, το ZipRecruiter μπορεί να εντοπίσει μοτίβα και συσχετίσεις μεταξύ της γλώσσας που χρησιμοποιείται στις περιγραφές θέσεων εργασίας και στα βιογραφικά των υποψηφίων. Αυτό επιτρέπει στην πλατφόρμα να προτείνει εξαιρετικά σχετικές ευκαιρίες απασχόλησης στους αναζητούντες εργασία και να παρουσιάζει στους εργοδότες εξειδικευμένους υποψηφίους που πληρούν τις συγκεκριμένες απαιτήσεις τους. Το αποτέλεσμα είναι μια πιο εκσυγχρονισμένη και αποτελεσματική διαδικασία πρόσληψης που ωφελεί και τα δύο μέρη (Tanberk et al., 2023).

Ανάλυση συναισθήματος εργαζομένων

Η εφαρμογή της Επεξεργασίας Φυσικής Γλώσσας (NLP) στους ανθρώπινους πόρους (HR) εκτείνεται πέρα από την πρόσληψη προσωπικού και περιλαμβάνει την ανάλυση του συναισθήματος των εργαζομένων. Η ανάλυση του συναισθήματος των εργαζομένων περιλαμβάνει τη χρήση εξελιγμένων εργαλείων NLP για την αξιολόγηση των αξιολογήσεων και των σχολίων των εργαζομένων, παρέχοντας πολύτιμες πληροφορίες για το συναίσθημα στο χώρο εργασίας. Η εφαρμογή αυτή είναι ιδιαίτερα σημαντική για την κατανόηση της συνολικής διάθεσης του εργατικού δυναμικού, τον εντοπισμό πιθανών προβλημάτων και την ενίσχυση της οργανωτικής κουλτούρας και της δέσμευσης των εργαζομένων (Fang & Zhan, 2015).

Η ανάλυση συναισθήματος εργαζομένων αξιοποιεί αλγορίθμους NLP για την επεξεργασία μεγάλου όγκου δεδομένων κειμένου που παράγονται από τους εργαζομένους, όπως κριτικές σε πλατφόρμες όπως η Glassdoor, φόρμες εσωτερικής ανατροφοδότησης και αναρτήσεις στα μέσα κοινωνικής δικτύωσης.

Αυτοί οι αλγόριθμοι αναλύουν τη γλώσσα που χρησιμοποιείται σε αυτά τα κείμενα για να ανιχνεύσουν συναισθήματα, απόψεις και στάσεις. Με την εξέταση μοτίβων και τάσεων στα δεδομένα, οι οργανισμοί μπορούν να αποκτήσουν βαθύτερη κατανόηση του τρόπου με τον οποίο οι εργαζόμενοι αισθάνονται για διάφορες πτυχές του εργασιακού τους περιβάλλοντος, συμπεριλαμβανομένων των πρακτικών διαχείρισης, των εταιρικών πολιτικών, της ισορροπίας μεταξύ επαγγελματικής και προσωπικής ζωής και των ευκαιριών εξέλιξης (Azhaguramyaa et al., 2022).

Ένα από τα βασικά εργαλεία που χρησιμοποιούνται για την ανάλυση συναισθήματος στο HR είναι το IBM Watson. Το IBM Watson χρησιμοποιεί προηγμένες τεχνικές NLP για την ανάλυση δεδομένων κειμένου και την εξαγωγή σημαντικών πληροφοριών. Στο πλαίσιο της ανάλυσης συναισθήματος των εργαζομένων, το IBM Watson μπορεί να επεξεργαστεί τα σχόλια των εργαζομένων για τον εντοπισμό θετικών, αρνητικών και ουδέτερων συναισθημάτων. Η ανάλυση αυτή βοηθά τους επαγγελματίες και τους διευθυντές ανθρώπινου δυναμικού να κατανοήσουν την επικρατούσα διάθεση εντός του οργανισμού και να εντοπίσουν τους τομείς που απαιτούν προσοχή ή βελτίωση. Για παράδειγμα, εάν η ανάλυση αποκαλύψει υψηλό επίπεδο αρνητικού συναισθήματος σχετικά με την ισορροπία μεταξύ επαγγελματικής και προσωπικής ζωής, ο οργανισμός μπορεί να λάβει προληπτικά μέτρα για την αντιμετώπιση αυτού του ζητήματος, εφαρμόζοντας πιο ευέλικτες ρυθμίσεις εργασίας ή παρέχοντας πρόσθετους πόρους υποστήριξης (Aslam et al., 2020).

Η Google έχει επίσης αναπτύξει πλατφόρμες που χρησιμοποιούν NLP για ανάλυση συναισθήματος, όπως το Google Cloud Natural Language API. Αυτό το εργαλείο μπορεί να αναλύσει τα σχόλια των εργαζομένων για τον εντοπισμό του συναισθήματος και την εξαγωγή βασικών θεμάτων. Με την ενσωμάτωση τέτοιων εργαλείων με τα συστήματα πληροφοριών ανθρώπινου δυναμικού, οι οργανισμοί μπορούν να παρακολουθούν συνεχώς το συναίσθημα των εργαζομένων και να λαμβάνουν αποφάσεις βάσει δεδομένων για τη βελτίωση του εργασιακού περιβάλλοντος. Για παράδειγμα, η ανάλυση συναισθήματος μπορεί να χρησιμοποιηθεί για την αξιολόγηση της αποτελεσματικότητας νέων πολιτικών ή πρωτοβουλιών, βοηθώντας τους οργανισμούς να επαναλαμβάνουν και να βελτιώνουν τις προσεγγίσεις τους με βάση την ανατροφοδότηση από τους εργαζόμενους σε πραγματικό χρόνο (Martinez-Gil et al., 2016).

Οι γνώσεις που αποκτώνται από την ανάλυση συναισθήματος των εργαζομένων έχουν διάφορες πρακτικές εφαρμογές. Πρώτον, επιτρέπουν στους οργανισμούς να αντιμετωπίζουν προληπτικά τις ανησυχίες των εργαζομένων πριν αυτές κλιμακωθούν σε μεγαλύτερα ζητήματα. Εντοπίζοντας έγκαιρα τους τομείς δυσaráεσκείας, οι επαγγελματίες του τομέα Ανθρώπινου Δυναμικού μπορούν να

εφαρμόσουν στοχευμένες παρεμβάσεις για τη βελτίωση του ηθικού και της δέσμευσης των εργαζομένων. Αυτή η προληπτική προσέγγιση όχι μόνο ενισχύει την ικανοποίηση των εργαζομένων αλλά και μειώνει τα ποσοστά κύκλου εργασιών, καθώς οι εργαζόμενοι είναι πιο πιθανό να παραμείνουν σε έναν οργανισμό που ακούει ενεργά και αντιμετωπίζει τις ανησυχίες τους (Azhaguramya et al., 2022).

Δεύτερον, η ανάλυση συναισθήματος υποστηρίζει την ανάπτυξη μιας θετικής οργανωτικής κουλτούρας. Η κατανόηση του τρόπου με τον οποίο οι εργαζόμενοι αντιλαμβάνονται τις διάφορες πτυχές του εργασιακού τους περιβάλλοντος επιτρέπει στους ηγέτες να ενισχύουν τις θετικές πρακτικές και να κάνουν τις απαραίτητες προσαρμογές σε τομείς που μπορεί να συμβάλλουν στο αρνητικό συναίσθημα. Για παράδειγμα, εάν η ανάλυση συναισθήματος αποκαλύπτει ότι οι εργαζόμενοι αισθάνονται υποτιμημένοι, ο οργανισμός μπορεί να εισαγάγει προγράμματα αναγνώρισης για την αναγνώριση και την επιβράβευση της συνεισφοράς των εργαζομένων. Αυτό προάγει μια κουλτούρα εκτίμησης και παρακίνησης, η οποία είναι απαραίτητη για τη διατήρηση υψηλών επιπέδων δέσμευσης και παραγωγικότητας των εργαζομένων (Fang & Zhan, 2015).

1.2.2. Οφέλη του NLP για την αναγνώριση δεξιοτήτων

Βελτιώσεις ακρίβειας

Η εφαρμογή της Επεξεργασίας Φυσικής Γλώσσας (NLP) στον τομέα της αναγνώρισης δεξιοτήτων έχει εγκαινιάσει μια νέα εποχή ακρίβειας και αποτελεσματικότητας στον εντοπισμό ικανοτήτων από αδόμητο κείμενο. Αυτή η τεχνολογική πρόοδος είναι ιδιαίτερα σημαντική δεδομένης της ολοένα και πιο δυναμικής και πολύπλοκης φύσης των αγορών εργασίας. Ένα από τα βασικά οφέλη της NLP στην αναγνώριση δεξιοτήτων είναι η αυξημένη ακρίβεια που επιφέρει στην εξαγωγή δεξιοτήτων από διάφορες μορφές κειμένου, όπως βιογραφικά σημειώματα, συνοδευτικές επιστολές και επαγγελματικά προφίλ. Αυτή η βελτίωση της ακρίβειας αποδίδεται σε μεγάλο βαθμό στην ανάπτυξη προηγμένων μοντέλων NLP, ιδίως του BERT (Bidirectional Encoder Representations from Transformers) και του GPT (Generative Pre-trained Transformer) (Wang et al., 2020).

Το BERT, που αναπτύχθηκε από την Google, αποτελεί ένα σημαντικό άλμα προς τα εμπρός στις δυνατότητες NLP λόγω της αμφίδρομης προσέγγισης εκπαίδευσης. Τα παραδοσιακά μοντέλα NLP συνήθως επεξεργάζονται κείμενο με διαδοχικό τρόπο, είτε από αριστερά προς τα δεξιά είτε από δεξιά προς τα αριστερά. Το BERT, ωστόσο, διαβάζει το κείμενο και προς τις δύο κατευθύνσεις ταυτόχρονα, επιτρέποντάς του να κατανοεί το πλαίσιο μιας λέξης με βάση τις

λέξεις που την περιβάλλουν. Αυτή η βαθιά κατανόηση των συμφραζομένων είναι ζωτικής σημασίας για τον ακριβή εντοπισμό δεξιοτήτων που ενσωματώνονται σε σύνθετες προτάσεις και ποικίλες μορφές. Για παράδειγμα, η BERT μπορεί να διακρίνει μεταξύ της "Python" ως γλώσσας προγραμματισμού και της "python" ως φιδιού, λαμβάνοντας υπόψη το πλαίσιο στο οποίο εμφανίζεται η λέξη. Αυτό το επίπεδο ακρίβειας είναι απαραίτητο για την ανάλυση βιογραφικών σημειωμάτων, όπου ο ίδιος όρος μπορεί να έχει διαφορετική σημασία ανάλογα με το πλαίσιο (Zhou et al., 2019).

Παρομοίως, τα μοντέλα GPT, ιδίως το GPT-3 που αναπτύχθηκε από την OpenAI, έχουν επιδείξει εξαιρετικές ικανότητες στην κατανόηση και παραγωγή κειμένου που μοιάζει με ανθρώπινο. Η εκτεταμένη εκπαίδευση του GPT-3 σε διάφορα σύνολα δεδομένων του επιτρέπει να κατανοεί και να εξάγει αποχρώσεις πληροφοριών από το κείμενο. Στο πλαίσιο της αναγνώρισης δεξιοτήτων, το GPT-3 μπορεί να αναλύσει προτάσεις για να εντοπίσει ρητές και έμμεσες αναφορές δεξιοτήτων, ακόμη και όταν αυτές δεν αναφέρονται με απλό τρόπο. Για παράδειγμα, ένα βιογραφικό σημείωμα μπορεί να περιγράψει ένα έργο στο οποίο ο υποψήφιος "ανέπτυξε έναν αλγόριθμο μηχανικής μάθησης για τη βελτιστοποίηση των διαδικασιών της αλυσίδας εφοδιασμού". Το GPT-3 μπορεί να συμπεράνει ότι αυτή η περιγραφή υποδηλώνει δεξιότητες στη μηχανική μάθηση, τη διαχείριση της εφοδιαστικής αλυσίδας και την ανάπτυξη αλγορίθμων. Αυτή η ικανότητα εξαγωγής συμπερασμάτων για δεξιότητες από λεπτομερείς περιγραφές ενισχύει την ακρίβεια και την πληρότητα των διαδικασιών εξαγωγής δεξιοτήτων (Wang et al., 2021).

Η ανάπτυξη αυτών των προηγμένων μοντέλων NLP σε συστήματα αναγνώρισης δεξιοτήτων επιφέρει πολλά απτά οφέλη. Πρώτον, η αυξημένη ακρίβεια στον εντοπισμό δεξιοτήτων διασφαλίζει ότι αναγνωρίζονται όλες οι σχετικές ικανότητες, γεγονός ιδιαίτερα πολύτιμο για υποψηφίους με διαφορετικές ή αντισυμβατικές πορείες σταδιοδρομίας. Η ακριβής εξαγωγή δεξιοτήτων επιτρέπει σε αυτούς τους υποψηφίους να προβάλλουν το πλήρες φάσμα των ικανοτήτων τους, αυξάνοντας τις πιθανότητες να ταιριάξουν με κατάλληλες ευκαιρίες απασχόλησης. Για τους εργοδότες, αυτή η ακρίβεια μεταφράζεται σε πιο αποτελεσματικές και αποδοτικές διαδικασίες πρόσληψης. Με τον ακριβή προσδιορισμό των δεξιοτήτων που απαιτούνται για μια θέση εργασίας και την αντιστοίχισή τους με τους κατάλληλους υποψηφίους, οι εργοδότες μπορούν να μειώσουν τον κίνδυνο αναντιστοιχιών στις προσλήψεις και να βελτιώσουν τη συνολική καταλληλότητα των θέσεων εργασίας (Yang et al., 2020).

Κέρδη αποδοτικότητας

Η ενσωμάτωση της Επεξεργασίας Φυσικής Γλώσσας (NLP) για την αναγνώριση δεξιοτήτων στην αγορά εργασίας έχει οδηγήσει σε σημαντική αύξηση της

αποδοτικότητα, ιδίως μειώνοντας σημαντικά τη χειρωνακτική προσπάθεια που απαιτείται για τον έλεγχο βιογραφικών σημειωμάτων. Ο μετασχηματισμός αυτός είναι καίριος σε μια εποχή όπου οι οργανισμοί λαμβάνουν συντριπτικό αριθμό αιτήσεων για κάθε θέση εργασίας, γεγονός που καθιστά αναγκαία μια πιο εκσυγχρονισμένη και αυτοματοποιημένη προσέγγιση για την αποτελεσματική διαχείριση της διαδικασίας πρόσληψης (Ali et al., 2022).

Ένα από τα πιο αξιοσημείωτα οφέλη της χρήσης της NLP στον έλεγχο βιογραφικών σημειωμάτων είναι η αυτοματοποίηση της ανάλυσης και της ανάλυσης μεγάλου όγκου βιογραφικών σημειωμάτων. Οι παραδοσιακές χειροκίνητες μέθοδοι διαλογής συνεπάγονται την σχολαστική ανάγνωση κάθε βιογραφικού από τους επαγγελματίες του ανθρώπινου δυναμικού για τον εντοπισμό των σχετικών δεξιοτήτων, προσόντων και εμπειριών. Αυτή η διαδικασία δεν είναι μόνο χρονοβόρα αλλά και επιρρεπής σε ανθρώπινα λάθη και ασυνέπειες. Αντίθετα, τα συστήματα που λειτουργούν με NLP μπορούν να επεξεργαστούν γρήγορα χιλιάδες βιογραφικά σημειώματα, εξάγοντας σχετικές πληροφορίες όπως δεξιότητες, εκπαιδευτικό υπόβαθρο και εργασιακή εμπειρία με μεγάλη ακρίβεια. Αυτά τα αυτοματοποιημένα συστήματα αξιοποιούν προηγμένους αλγόριθμους για τη σάρωση και την ερμηνεία του κειμένου μέσα στα βιογραφικά σημειώματα, διασφαλίζοντας ότι τα βασικά προσόντα εντοπίζονται και επισημαίνονται αποτελεσματικά (Anusha et al., 2023).

Η ανάπτυξη αυτοματοποιημένων συστημάτων διαλογής βιογραφικών σημειωμάτων έχει γίνει παράδειγμα από πλατφόρμες όπως το LinkedIn και το Indeed, οι οποίες χρησιμοποιούν NLP για να ενισχύσουν τις διαδικασίες πρόσληψής τους. Για παράδειγμα, το εργαλείο Recruiter του LinkedIn χρησιμοποιεί εξελιγμένους αλγόριθμους NLP για την ανάλυση και την αντιστοίχιση βιογραφικών σημειωμάτων με τις περιγραφές θέσεων εργασίας, μειώνοντας σημαντικά τη χειροκίνητη προσπάθεια που απαιτείται για τον έλεγχο των υποψηφίων. Με την αυτοματοποίηση αυτού του αρχικού σταδίου, το LinkedIn δίνει τη δυνατότητα στους υπεύθυνους προσλήψεων να επικεντρώσουν τις προσπάθειές τους σε πιο στρατηγικά καθήκοντα, όπως η συνέντευξη και η επικοινωνία με δυνητικούς υποψηφίους. Αυτή η μετατόπιση όχι μόνο επιταχύνει τη διαδικασία πρόσληψης αλλά και ενισχύει τη συνολική αποτελεσματικότητα των στρατηγικών απόκτησης ταλέντων (Harsha et al., 2022).

Η Indeed, μια άλλη εξέχουσα πλατφόρμα αναζήτησης εργασίας, χρησιμοποιεί NLP για τη διαχείριση και επεξεργασία του τεράστιου αριθμού βιογραφικών που υποβάλλονται καθημερινά. Το σύστημα της Indeed σαρώνει τα βιογραφικά σημειώματα για να εντοπίσει λέξεις-κλειδιά και φράσεις που ταιριάζουν με τις απαιτήσεις των αγγελιών εργασίας. Αυτή η αυτοματοποιημένη προσέγγιση διασφαλίζει ότι μόνο οι πιο σχετικοί υποψήφιοι μπαίνουν στη λίστα,

εξοικονομώντας έτσι σημαντικό χρόνο για τους επαγγελματίες του τομέα Ανθρώπινου Δυναμικού. Η ικανότητα ταχείας διαχείρισης μεγάλου όγκου βιογραφικών σημειωμάτων επιτρέπει στους οργανισμούς να διατηρούν ανταγωνιστικό πλεονέκτημα, καλύπτοντας θέσεις πιο γρήγορα και αποτελεσματικά (Ni et al., 2015).

Τα κέρδη αποδοτικότητας από τη NLP στον έλεγχο βιογραφικών σημειωμάτων επεκτείνονται πέρα από την απλή ταχύτητα και ακρίβεια. Μεταφράζονται επίσης σε σημαντική εξοικονόμηση κόστους για τους οργανισμούς. Ο χειροκίνητος έλεγχος βιογραφικών απαιτεί σημαντικούς ανθρώπινους πόρους, γεγονός που συχνά καθιστά αναγκαία την ύπαρξη μεγάλων ομάδων ανθρώπινου δυναμικού ή την εξωτερική ανάθεση σε γραφεία προσλήψεων. Τα αυτοματοποιημένα συστήματα, ωστόσο, μπορούν να λειτουργούν συνεχώς και να χειρίζονται εκτεταμένα σύνολα δεδομένων χωρίς πρόσθετη ανθρώπινη παρέμβαση. Αυτή η επεκτασιμότητα σημαίνει ότι οι οργανισμοί μπορούν να διαχειρίζονται κυμαινόμενους όγκους αιτήσεων εργασίας χωρίς να επιβαρύνονται με ανάλογες αυξήσεις στο κόστος προσωπικού. Επιπλέον, μειώνοντας την εξάρτηση από εξωτερικές υπηρεσίες προσλήψεων, οι εταιρείες μπορούν να κατανέμουν τους πόρους αποτελεσματικότερα, εστιάζοντας σε άλλους κρίσιμους τομείς της διαχείρισης ανθρώπινου δυναμικού και της οργανωτικής ανάπτυξης (Anusha et al., 2023).

Συνέπεια και επεκτασιμότητα

Η υιοθέτηση της Επεξεργασίας Φυσικής Γλώσσας (NLP) για την αναγνώριση δεξιοτήτων στην αγορά εργασίας προσφέρει σημαντικά οφέλη όσον αφορά τη συνέπεια και την επεκτασιμότητα. Τα πλεονεκτήματα αυτά είναι ιδιαίτερα σημαντικά στο πλαίσιο του ελέγχου βιογραφικών σημειωμάτων και της αξιολόγησης υποψηφίων, όπου η ικανότητα εφαρμογής ομοίμορφων κριτηρίων και αποτελεσματικής διαχείρισης διαφορετικών όγκων αιτήσεων είναι υψίστης σημασίας. Οι τεχνολογίες NLP, ιδίως όταν αναπτύσσονται μέσω υπηρεσιών που βασίζονται στο υπολογιστικό νέφος, διασφαλίζουν ότι η εξαγωγή δεξιοτήτων και η αξιολόγηση υποψηφίων διεξάγονται με υψηλό βαθμό συνέπειας, ανεξάρτητα από την κλίμακα των εργασιών (Pallas et al., 2020).

Η συνέπεια στην εφαρμογή των κριτηρίων εξαγωγής δεξιοτήτων είναι ένα κρίσιμο πλεονέκτημα που παρέχουν οι τεχνολογίες NLP. Οι παραδοσιακές μέθοδοι διαλογής βιογραφικών σημειωμάτων, οι οποίες βασίζονται σε μεγάλο βαθμό στην ανθρώπινη κρίση, είναι επιρρεπείς σε ασυνέπειες και προκαταλήψεις. Διαφορετικοί υπεύθυνοι προσλήψεων ενδέχεται να ερμηνεύουν τις ίδιες πληροφορίες με διαφορετικούς τρόπους, οδηγώντας σε αποκλίσεις στην αξιολόγηση των υποψηφίων (Stokes et al., 2015). Τα συστήματα NLP, ωστόσο, χρησιμοποιούν προκαθορισμένους αλγόριθμους για την ανάλυση βιογραφικών

σημειωμάτων, διασφαλίζοντας ότι τα ίδια κριτήρια εφαρμόζονται ομοιόμορφα σε όλα τα έγγραφα. Αυτή η συνεπής εφαρμογή των προτύπων αξιολόγησης είναι απαραίτητη για τη διατήρηση της δικαιοσύνης και της αντικειμενικότητας στη διαδικασία πρόσληψης. Διασφαλίζει ότι όλοι οι υποψήφιοι αξιολογούνται με βάση τις δεξιότητες και τα προσόντα τους και όχι με βάση υποκειμενικούς παράγοντες, προωθώντας έτσι μια πιο δίκαιη διαδικασία πρόσληψης (Chen et al., 2014).

Οι υπηρεσίες NLP που βασίζονται στο νέφος ενισχύουν περαιτέρω αυτή τη συνοχή προσφέροντας κεντρικές πλατφόρμες όπου τα μοντέλα εξαγωγής δεξιοτήτων μπορούν να ενημερώνονται και να βελτιώνονται συνεχώς. Αυτές οι πλατφόρμες επιτρέπουν στους οργανισμούς να αναπτύσσουν ομοιόμορφα μοντέλα NLP σε όλα τα υποκαταστήματα και τα τμήματα, διασφαλίζοντας ότι τα ίδια υψηλά πρότυπα διαλογής βιογραφικών σημειωμάτων διατηρούνται σε ολόκληρο τον οργανισμό. Για παράδειγμα, οι εταιρείες που χρησιμοποιούν υπηρεσίες που βασίζονται στο cloud, όπως το Google Cloud Natural Language ή το Amazon Comprehend, επωφελούνται από τις τακτικές ενημερώσεις και βελτιώσεις των αλγορίθμων NLP, οι οποίες ενσωματώνονται απρόσκοπτα σε όλες τις διεπαφές χρήστη. Αυτή η κεντρική προσέγγιση διασφαλίζει ότι η εξαγωγή δεξιοτήτων παραμένει συνεπής, ακριβής και ενημερωμένη, ανεξάρτητα από τις αλλαγές στην αγορά εργασίας ή την εισαγωγή νέων ορολογιών (Liang et al., 2024).

Η επεκτασιμότητα είναι ένα άλλο σημαντικό πλεονέκτημα της χρήσης του NLP για την αναγνώριση δεξιοτήτων. Καθώς οι οργανισμοί αναπτύσσονται και οι ανάγκες πρόσληψής τους διευρύνονται, ο όγκος των βιογραφικών σημειωμάτων και των αιτήσεων εργασίας μπορεί να αυξηθεί εκθετικά. Οι χειροκίνητες μέθοδοι διαλογής καθίστανται ανέφικτες σε τέτοιες κλίμακες, οδηγώντας σε συμφόρηση και αναποτελεσματικότητα. Τα συστήματα NLP, ειδικά εκείνα που φιλοξενούνται σε πλατφόρμες cloud, μπορούν εύκολα να κλιμακωθούν για να φιλοξενήσουν ποικίλους όγκους δεδομένων χωρίς συμβιβασμούς στην απόδοση ή την ακρίβεια. Αυτά τα συστήματα έχουν σχεδιαστεί για να χειρίζονται μεγάλα σύνολα δεδομένων, επεξεργαζόμενα χιλιάδες βιογραφικά σημειώματα ταυτόχρονα και αποτελεσματικά (Pallas et al., 2020).

1.3. Εργασίες NLP σχετικές με την ανάλυση της αγοράς εργασίας

1.3.1. Ταξινόμηση κειμένου

Ορισμός και διαδικασία

Η ταξινόμηση κειμένου είναι μια θεμελιώδης εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει την ανάθεση προκαθορισμένων κατηγοριών σε έγγραφα κειμένου. Η διαδικασία αυτή είναι ζωτικής σημασίας για την οργάνωση και τη διαχείριση μεγάλου όγκου δεδομένων κειμένου, επιτρέποντας την αποτελεσματική ανάκτηση και ανάλυση. Στο πλαίσιο της αγοράς εργασίας, η ταξινόμηση κειμένου διαδραματίζει καθοριστικό ρόλο στην ταξινόμηση βιογραφικών σημειωμάτων, περιγραφών θέσεων εργασίας και άλλων εγγράφων που σχετίζονται με την απασχόληση σε συγκεκριμένες κατηγορίες, όπως τομείς του κλάδου, ρόλοι εργασίας και σύνολα δεξιοτήτων. Αυτή η συστηματική οργάνωση βοηθά τους εργοδότες στον εντοπισμό των κατάλληλων υποψηφίων και επιτρέπει στους αιτούντες εργασία να αντιστοιχίζονται με τις κατάλληλες ευκαιρίες απασχόλησης (Goller et al., 2000).

Η διαδικασία της ταξινόμησης κειμένου ξεκινά με τη συλλογή ενός μεγάλου σώματος δεδομένων κειμένου. Στη συνέχεια, τα δεδομένα αυτά υποβάλλονται σε προεπεξεργασία για την απομάκρυνση του θορύβου και των άσχετων πληροφοριών, εξασφαλίζοντας ότι το κείμενο έχει κατάλληλη μορφή για ανάλυση. Τα στάδια προεπεξεργασίας περιλαμβάνουν συνήθως tokenization, όπου το κείμενο διαιρείται σε μεμονωμένες λέξεις ή φράσεις, stemming και lemmatization, τα οποία μειώνουν τις λέξεις στις ριζικές τους μορφές, και την αφαίρεση των stop words, όπως "και", "το" και "του", οι οποίες δεν συνεισφέρουν σημαντικό νόημα. Αυτά τα καθαρισμένα και δομημένα δεδομένα κειμένου χρησιμοποιούνται στη συνέχεια για την εκπαίδευση αλγορίθμων ταξινόμησης (HaCohen-Kerner et al., 2020).

Διάφοροι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την ταξινόμηση κειμένου, από παραδοσιακές μεθόδους όπως ο Naive Bayes και οι μηχανές διανυσμάτων υποστήριξης (SVM) έως πιο προηγμένες τεχνικές όπως τα νευρωνικά δίκτυα και τα μοντέλα βαθιάς μάθησης. Αυτοί οι αλγόριθμοι μαθαίνουν να αναγνωρίζουν μοτίβα και σχέσεις μέσα στα δεδομένα κειμένου, επιτρέποντάς τους να προβλέπουν την κατάλληλη κατηγορία για νέα, αθέατα έγγραφα. Για παράδειγμα, ένας ταξινομητής SVM μπορεί να εκπαιδευτεί σε ένα σύνολο δεδομένων με βιογραφικά σημειωμένα με διαφορετικούς κλάδους. Μαθαίνοντας τα διακριτικά χαρακτηριστικά των βιογραφικών σημειωμάτων σε τομείς όπως η πληροφορική, τα οικονομικά και η υγειονομική περίθαλψη, ο ταξινομητής μπορεί να κατηγοριοποιήσει με ακρίβεια νέα βιογραφικά σημειώματα με βάση το περιεχόμενό τους (Taha & Abuelsoud, 2022).

Μία από τις βασικές εφαρμογές της ταξινόμησης κειμένου στην αγορά εργασίας είναι η κατηγοριοποίηση βιογραφικών σημειωμάτων σε συγκεκριμένα πεδία. Η διαδικασία αυτή περιλαμβάνει την ανάλυση του κειμένου των βιογραφικών σημειωμάτων για τον εντοπισμό του κλάδου ή του ρόλου εργασίας που ταιριάζει

καλύτερα με την εμπειρία και τις δεξιότητες του υποψηφίου. Για παράδειγμα, ένα βιογραφικό σημείωμα που αναφέρει γλώσσες προγραμματισμού όπως η Python και η Java, frameworks όπως το Django και το Spring και εμπειρία σε έργα ανάπτυξης λογισμικού θα μπορούσε να ταξινομηθεί στον τομέα της πληροφορικής. Παρομοίως, ένα βιογραφικό σημείωμα που τονίζει τις γνώσεις χρηματοοικονομικής μοντελοποίησης, διαχείρισης κινδύνων και κανονιστικής συμμόρφωσης θα κατατασσόταν στον τομέα των χρηματοοικονομικών. Με την αυτόματη ταξινόμηση των βιογραφικών σημειωμάτων σε αυτές τις προκαθορισμένες κατηγορίες, οι εργοδότες μπορούν να διαχειριστούν αποτελεσματικότερα τη διαδικασία πρόσληψης, εντοπίζοντας γρήγορα τους υποψηφίους που πληρούν τις συγκεκριμένες απαιτήσεις μιας θέσης εργασίας (Goller et al., 2000).

Μια άλλη σημαντική εφαρμογή είναι η ταξινόμηση περιγραφών θέσεων εργασίας. Οι εργοδότες πρέπει συχνά να κατηγοριοποιούν τις προκηρύξεις θέσεων εργασίας για να διασφαλίσουν ότι θα φτάσουν στο σωστό κοινό και να βελτιώσουν τη διαδικασία υποβολής αιτήσεων. Οι αλγόριθμοι ταξινόμησης κειμένου μπορούν να αναλύσουν τη γλώσσα και την ορολογία που χρησιμοποιείται στις περιγραφές θέσεων εργασίας για να προσδιορίσουν τον κατάλληλο κλάδο, ρόλο και σύνολο δεξιοτήτων. Για παράδειγμα, μια θέση εργασίας που απαιτεί εξειδίκευση στα συστήματα πληροφορικής υγείας, στη διαχείριση δεδομένων ασθενών και στη συμμόρφωση με τους κανονισμούς της υγειονομικής περίθαλψης θα μπορούσε να ταξινομηθεί στον τομέα της πληροφορικής υγείας. Αυτή η κατηγοριοποίηση βοηθά τους αναζητούντες εργασία να βρίσκουν ευκολότερα τις σχετικές ευκαιρίες απασχόλησης και επιτρέπει στους πίνακες εργασίας και τις πλατφόρμες προσλήψεων να οργανώνουν αποτελεσματικότερα τις καταχωρίσεις θέσεων εργασίας (HaCohen-Kerner et al., 2020).

Τεχνικές και μοντέλα

Η ταξινόμηση κειμένου, μια κρίσιμη εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP), χρησιμοποιεί διάφορες τεχνικές και μοντέλα για την κατηγοριοποίηση δεδομένων κειμένου σε προκαθορισμένες ομάδες. Η ικανότητα αυτή είναι ιδιαίτερα σημαντική στην ανάλυση της αγοράς εργασίας, όπου η αποτελεσματική επεξεργασία και ταξινόμηση βιογραφικών σημειωμάτων, περιγραφών θέσεων εργασίας και άλλων εγγράφων που σχετίζονται με την απασχόληση είναι απαραίτητη. Οι προσεγγίσεις μηχανικής μάθησης, συμπεριλαμβανομένων των Μηχανών Διανυσμάτων Υποστήριξης (SVM), των Naïve Bayes και των νευρωνικών δικτύων, διαδραματίζουν καθοριστικό ρόλο σε αυτό το πλαίσιο, παρέχοντας ισχυρά πλαίσια για την ακριβή ταξινόμηση και ανάλυση μεγάλων όγκων δεδομένων κειμένου (Pappas & Popescu-Belis, 2017).

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι μια δημοφιλής επιλογή για την ταξινόμηση κειμένου λόγω της αποτελεσματικότητάς τους σε χώρους υψηλών διαστάσεων. Οι SVM λειτουργούν με την εύρεση του υπερεπιπέδου που διαχωρίζει καλύτερα τα σημεία δεδομένων σε διαφορετικές κλάσεις (Alfaro et al., 2023). Στο πλαίσιο της ταξινόμησης κειμένου, οι SVM μπορούν να εκπαιδευτούν σε επισημασμένα σύνολα δεδομένων βιογραφικών σημειωμάτων ή περιγραφών θέσεων εργασίας. Μαθαίνοντας τα διακριτικά χαρακτηριστικά των διαφόρων κατηγοριών, όπως οι κλάδοι ή οι ρόλοι εργασίας, οι SVM μπορούν να ταξινομήσουν με ακρίβεια νέα, αθέατα έγγραφα. Για παράδειγμα, ένας SVM μπορεί να εκπαιδευτεί για να αναγνωρίζει βιογραφικά που σχετίζονται με την πληροφορική, εντοπίζοντας λέξεις-κλειδιά και φράσεις που σχετίζονται με τον προγραμματισμό, την ανάπτυξη λογισμικού και τη διαχείριση τεχνολογίας. Όταν παρουσιάζεται ένα νέο βιογραφικό σημείωμα, ο SVM μπορεί να το ταξινομήσει στην κατάλληλη κατηγορία με βάση αυτά τα μαθημένα χαρακτηριστικά (Maia et al., 2021).

Οι ταξινομητές Naive Bayes, μια άλλη ευρέως χρησιμοποιούμενη τεχνική, βασίζονται σε πιθανοτικά μοντέλα που εφαρμόζουν το θεώρημα του Bayes με την υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών. Παρά την απλουστευτική αυτή υπόθεση, οι ταξινομητές Naive Bayes είναι ιδιαίτερα αποτελεσματικοί για εργασίες ταξινόμησης κειμένου, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων. Στην αγορά εργασίας, ο Naive Bayes μπορεί να χρησιμοποιηθεί για την ταξινόμηση περιγραφών θέσεων εργασίας αναλύοντας τη συχνότητα συγκεκριμένων όρων. Για παράδειγμα, οι προκηρξίες θέσεων εργασίας που αναφέρουν συχνά όρους όπως "χρηματοοικονομική ανάλυση", "επενδυτικές στρατηγικές" και "διαχείριση κινδύνου" μπορούν να ταξινομηθούν στην κατηγορία των χρηματοοικονομικών. Η πιθανολογική φύση του Naive Bayes του επιτρέπει να χειρίζεται τη μεταβλητότητα στη χρήση της γλώσσας, καθιστώντας το ένα ισχυρό εργαλείο για την ταξινόμηση ποικίλου κειμενικού περιεχομένου.

Τα νευρωνικά δίκτυα, ιδίως τα μοντέλα βαθιάς μάθησης, έχουν προωθήσει σημαντικά τον τομέα της ταξινόμησης κειμένου. Τα συνεπαγωγικά νευρωνικά δίκτυα (CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) είναι δύο εξέχουσες αρχιτεκτονικές που χρησιμοποιούνται για το σκοπό αυτό. Τα CNNs, που αναπτύχθηκαν αρχικά για την επεξεργασία εικόνας, έχουν προσαρμοστεί για την ταξινόμηση κειμένου αντιμετωπίζοντας τα δεδομένα κειμένου ως μονοδιάστατη ακολουθία. Τα CNNs χρησιμοποιούν συνελκτικά στρώματα για την αυτόματη ανίχνευση και εκμάθηση μοτίβων στο κείμενο, όπως συνδυασμοί λέξεων και φράσεις που είναι ενδεικτικές συγκεκριμένων κατηγοριών. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για τον εντοπισμό αποχρώσεων που μπορεί

να μην ληφθούν υπόψη από τα παραδοσιακά μοντέλα. Στην ανάλυση της αγοράς εργασίας, τα CNN μπορούν να ταξινομήσουν βιογραφικά σημειώματα αναγνωρίζοντας σύνθετα σύνολα δεξιοτήτων και εμπειριών που συσχετίζονται με συγκεκριμένους ρόλους εργασίας (Pappas & Popescu-Belis, 2017).

Τα RNN, σχεδιασμένα για να χειρίζονται διαδοχικά δεδομένα, είναι ιδιαίτερα κατάλληλα για εργασίες ταξινόμησης κειμένου που περιλαμβάνουν συμφραζόμενες και χρονικές εξαρτήσεις. Τα RNN και οι παραλλαγές τους, όπως τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM), διατηρούν μνήμη των προηγούμενων λέξεων σε μια ακολουθία, επιτρέποντάς τους να κατανοούν το πλαίσιο και το νόημα των προτάσεων πιο αποτελεσματικά. Αυτή η ικανότητα είναι ζωτικής σημασίας για την ακριβή ταξινόμηση κειμένων όπου το πλαίσιο παίζει σημαντικό ρόλο. Για παράδειγμα, ένα RNN μπορεί να διακρίνει μεταξύ ενός βιογραφικού σημειώματος που περιγράφει λεπτομερώς την εμπειρία στη "διαχείριση έργων" έναντι της "έρευνας έργων", ακόμη και αν το περιβάλλον κείμενο είναι παρόμοιο. Διατηρώντας πληροφορίες σχετικά με το πλαίσιο, τα RNN παρέχουν μια πιο εκλεπτυσμένη και ακριβή ταξινόμηση (Alfaro et al., 2023).

1.3.2. Αναγνώριση ονομαστικών οντοτήτων (NER)

Ορισμός και ρόλος

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι μια κρίσιμη εργασία στο πλαίσιο της επεξεργασίας φυσικής γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την ταξινόμηση ονομαστικών οντοτήτων σε κείμενο σε προκαθορισμένες κατηγορίες, όπως ονόματα ανθρώπων, οργανισμών, ημερομηνιών, τοποθεσιών και άλλα. Η διαδικασία αυτή είναι θεμελιώδης για την εξαγωγή δομημένων πληροφοριών από αδόμητο κείμενο, μετατρέποντας έτσι τα ακατέργαστα δεδομένα κειμένου σε μορφή που αναλύεται ευκολότερα και αξιοποιείται για διάφορες εφαρμογές. Στο πλαίσιο της ανάλυσης της αγοράς εργασίας, η NER διαδραματίζει σημαντικό ρόλο επιτρέποντας την αυτοματοποιημένη εξαγωγή σχετικών οντοτήτων από έγγραφα όπως βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας, διευκολύνοντας την αποτελεσματικότερη και ακριβέστερη επεξεργασία δεδομένων και τη λήψη αποφάσεων (Prokofyev et al., 2014).

Ο πρωταρχικός στόχος της NER είναι ο εντοπισμός και η κατηγοριοποίηση των οντοτήτων που αναφέρονται στο κείμενο. Για παράδειγμα, κατά την επεξεργασία ενός βιογραφικού σημειώματος, ένα σύστημα NER θα εντοπίσει το όνομα του υποψηφίου, τα εκπαιδευτικά ιδρύματα στα οποία φοίτησε, τους προηγούμενους εργοδότες, τους τίτλους εργασίας και τις ημερομηνίες απασχόλησης. Με τον τρόπο αυτό, το NER μετατρέπει τα αδόμητα δεδομένα βιογραφικού σε δομημένες

πληροφορίες, οι οποίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τον εμπλουτισμό βάσεων δεδομένων, την αντιστοίχιση υποψηφίων με κενές θέσεις εργασίας και τη δημιουργία συνοπτικών αναφορών. Αυτή η αυτοματοποίηση όχι μόνο εξοικονομεί σημαντικό χρόνο και προσπάθεια για τους επαγγελματίες του ανθρώπινου δυναμικού, αλλά εξασφαλίζει επίσης υψηλότερο βαθμό ακρίβειας και συνέπειας στο χειρισμό των δεδομένων (Prabhakar et al., 2014).

Η διαδικασία NER περιλαμβάνει διάφορα στάδια. Αρχικά, το κείμενο υποβάλλεται σε προεπεξεργασία, η οποία περιλαμβάνει την κατανομή των συμβόλων (tokenization), όπου το κείμενο αναλύεται σε μεμονωμένες λέξεις ή σημεία (tokens), και την επισήμανση των μερών του λόγου (part-of-speech tagging), η οποία προσδιορίζει τα γραμματικά μέρη του λόγου κάθε σημείου. Στη συνέχεια, το σύστημα NER εφαρμόζει διάφορους αλγορίθμους και μοντέλα για την αναγνώριση και ταξινόμηση οντοτήτων. Οι παραδοσιακές μέθοδοι NER περιλάμβαναν προσεγγίσεις που βασίζονται σε κανόνες και στατιστικά μοντέλα, αλλά οι πρόσφατες εξελίξεις έχουν δει την υιοθέτηση τεχνικών μηχανικής μάθησης και μοντέλων βαθιάς μάθησης, τα οποία προσφέρουν ανώτερες επιδόσεις και προσαρμοστικότητα (Su et al., 2022).

Τα συστήματα NER που βασίζονται στη μηχανική μάθηση εκπαιδεύονται σε σχολιασμένα σύνολα δεδομένων, όπου οι οντότητες στο κείμενο επισημαίνονται με τις αντίστοιχες κατηγορίες τους. Μέσω αυτής της διαδικασίας εκπαίδευσης, το σύστημα μαθαίνει να αναγνωρίζει μοτίβα και χαρακτηριστικά που διακρίνουν διαφορετικούς τύπους οντοτήτων. Για παράδειγμα, το πλαίσιο στο οποίο εμφανίζεται μια λέξη με κεφαλαία γράμματα μπορεί να βοηθήσει το σύστημα να προσδιορίσει αν πρόκειται για το όνομα ενός ατόμου ή ενός οργανισμού. Αφού εκπαιδευτεί, το σύστημα NER μπορεί να εφαρμοστεί σε νέο, άρατο κείμενο για να εντοπίσει και να ταξινομήσει οντότητες με υψηλή ακρίβεια (Manjula & Shridhar, 2019).

Στο πλαίσιο της ανάλυσης της αγοράς εργασίας, η NER είναι ιδιαίτερα πολύτιμη για την ανάλυση βιογραφικών σημειωμάτων. Εξάγοντας αυτόματα οντότητες όπως ονόματα, τίτλους θέσεων εργασίας, εταιρείες, ημερομηνίες και εκπαιδευτικά ιδρύματα, τα συστήματα NER μπορούν να δημιουργήσουν δομημένα προφίλ υποψηφίων. Αυτά τα δομημένα δεδομένα μπορούν στη συνέχεια να αναζητηθούν εύκολα, να φιλτραριστούν και να αντιστοιχιστούν με περιγραφές θέσεων εργασίας, απλοποιώντας τη διαδικασία πρόσληψης. Για παράδειγμα, ένα σύστημα NER μπορεί να εντοπίσει γρήγορα όλα τα βιογραφικά που αναφέρουν ως τίτλο εργασίας "Data Scientist" και ως δεξιότητα "Python", βοηθώντας έτσι τους υπεύθυνους προσλήψεων να εστιάσουν την προσοχή τους στους πιο κατάλληλους υποψηφίους (Prokofyev et al., 2014).

Επιπλέον, η NER συμβάλλει στην ανάλυση περιγραφών θέσεων εργασίας για την εξαγωγή βασικών οντοτήτων που καθορίζουν τις απαιτήσεις και τα χαρακτηριστικά των θέσεων εργασίας. Οντότητες όπως οι απαιτούμενες δεξιότητες, οι τίτλοι θέσεων εργασίας, τα ονόματα εταιρειών και οι τοποθεσίες μπορούν να εξαχθούν και να χρησιμοποιηθούν για την κατηγοριοποίηση και την οργάνωση των καταχωρίσεων θέσεων εργασίας. Αυτό επιτρέπει στους αναζητούντες εργασία να βρίσκουν πιο αποτελεσματικά τις σχετικές ευκαιρίες και επιτρέπει στους εργοδότες να προσεγγίζουν ένα πιο στοχευμένο κοινό (Prabhakar et al., 2014).

Τεχνικές και μοντέλα

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι μια βασική εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την ταξινόμηση οντοτήτων σε κείμενο σε προκαθορισμένες κατηγορίες, όπως ονόματα, ημερομηνίες, οργανισμούς και τοποθεσίες. Οι τεχνικές NER έχουν εξελιχθεί σημαντικά, από προσεγγίσεις που βασίζονται σε κανόνες έως εξελιγμένα μοντέλα μηχανικής μάθησης, καθένα από τις οποίες προσφέρει μοναδικά πλεονεκτήματα και εφαρμογές σε διάφορα πλαίσια, συμπεριλαμβανομένης της ανάλυσης της αγοράς εργασίας (Yu et al., 2020).

Οι μέθοδοι βασισμένες σε κανόνες ήταν από τις πρώτες προσεγγίσεις στην ΕΠΑ. Τα συστήματα αυτά βασίζονται σε ένα σύνολο χειροποίητων κανόνων και μοτίβων για τον εντοπισμό οντοτήτων στο κείμενο. Για παράδειγμα, ένας κανόνας μπορεί να ορίζει ότι οι λέξεις με κεφαλαία γράμματα που ακολουθούν ορισμένες λέξεις-κλειδιά όπως "Mr." ή "Dr." είναι πιθανό να είναι ονόματα. Ενώ τα συστήματα που βασίζονται σε κανόνες μπορεί να είναι εξαιρετικά ακριβή για συγκεκριμένες εργασίες και είναι εύκολα στην ερμηνεία, συχνά είναι περιορισμένα σε πεδίο εφαρμογής και ευελιξία. Δυσκολεύονται να αντιμετωπίσουν τη μεταβλητότητα και την ασάφεια της φυσικής γλώσσας, όπως η αντιμετώπιση νέων τύπων οντοτήτων ή διαφορετικών στυλ γραφής. Επομένως, ενώ είναι χρήσιμα σε ελεγχόμενα περιβάλλοντα, τα συστήματα NER που βασίζονται σε κανόνες είναι λιγότερο αποτελεσματικά σε δυναμικές, πραγματικές εφαρμογές όπου η χρήση της γλώσσας μπορεί να είναι απρόβλεπτη (Lothritz et al., 2020).

Για να αντιμετωπιστούν οι περιορισμοί των συστημάτων που βασίζονται σε κανόνες, αναπτύχθηκαν στατιστικές μέθοδοι όπως τα Conditional Random Fields (CRF). Το CRF είναι ένας τύπος πιθανοτικού γραφικού μοντέλου που υπερέχει σε εργασίες επίσημης ακολουθιών, καθιστώντας το κατάλληλο για NER. Σε ένα μοντέλο CRF, η πιθανότητα μιας ακολουθίας ετικετών (π.χ. ετικέτες οντοτήτων) μοντελοποιείται δεδομένης της ακολουθίας των παρατηρούμενων δεδομένων (π.χ. λέξεις σε ένα κείμενο). Οι CRF λαμβάνουν υπόψη το πλαίσιο μιας λέξης μέσα

σε μια πρόταση, επιτρέποντας στο μοντέλο να καταγράφει εξαρτήσεις μεταξύ ετικετών, γεγονός που ενισχύει την ικανότητά του να αναγνωρίζει με ακρίβεια οντότητες. Για παράδειγμα, σε ένα βιογραφικό σημείωμα, οι CRF μπορούν να διακρίνουν αποτελεσματικά μεταξύ του ονόματος ενός ατόμου και της επωνυμίας μιας εταιρείας με βάση τις περιβάλλουσες λέξεις και τα συμφραζόμενα (Hu et al., 2022).

Η έλευση της βαθιάς μάθησης μεταμόρφωσε περαιτέρω την ΕΑΑ με την εισαγωγή μοντέλων όπως τα δίκτυα αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (BiLSTM). Τα μοντέλα BiLSTM είναι ένας τύπος επαναλαμβανόμενου νευρωνικού δικτύου (RNN) που επεξεργάζεται κείμενο τόσο προς τα εμπρός όσο και προς τα πίσω. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο μοντέλο να συλλάβει πληροφορίες σχετικά με το πλαίσιο τόσο από τις παρελθούσες όσο και από τις μελλοντικές λέξεις, οδηγώντας σε ακριβέστερη αναγνώριση οντοτήτων. Τα BiLSTM είναι ιδιαίτερα αποτελεσματικά στο χειρισμό των εξαρτήσεων μεγάλης εμβέλειας στο κείμενο, κάτι που είναι ζωτικής σημασίας για την κατανόηση του πλαισίου στο οποίο εμφανίζεται μια οντότητα. Για παράδειγμα, σε μια περιγραφή εργασίας, ένα BiLSTM μπορεί να αναγνωρίσει με ακρίβεια την "Python" ως γλώσσα προγραμματισμού λαμβάνοντας υπόψη ολόκληρη τη δομή και το πλαίσιο της πρότασης, αντί να βασίζεται αποκλειστικά σε τοπικά μοτίβα λέξεων (Liu et al., 2020).

Οι μετασχηματιστές αντιπροσωπεύουν την τελευταία εξέλιξη στον τομέα της ΝΕΡ, προσφέροντας πρωτοφανή ακρίβεια και ευελιξία στην εργασία. Μοντέλα όπως το BERT (Bidirectional Encoder Representations from Transformers) και το GPT (Generative Pre-trained Transformer) έχουν θέσει νέα σημεία αναφοράς για την απόδοση της ΝΕΡ. Το BERT, ειδικότερα, αξιοποιεί μια αρχιτεκτονική αμφίδρομου μετασχηματιστή για να καταγράφει το περιεχόμενο και από τις δύο κατευθύνσεις μιας ακολουθίας κειμένου ταυτόχρονα. Αυτή η βαθιά κατανόηση των συμφραζομένων επιτρέπει στον BERT να αναγνωρίζει οντότητες με υψηλή ακρίβεια, ακόμη και σε σύνθετες και διαφοροποιημένες προτάσεις. Για παράδειγμα, η BERT μπορεί να διακρίνει μεταξύ της "Apple" ως φρούτο και της "Apple" ως εταιρεία με βάση τα στοιχεία συμφραζομένων από ολόκληρη την πρόταση (Lothritz et al., 2020).

1.3.3. Εξαγωγή λέξεων-κλειδιών

Ορισμός και διαδικασία

Η εξαγωγή λέξεων-κλειδιών είναι μια θεμελιώδης εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την εξαγωγή σημαντικών λέξεων και φράσεων από ένα κείμενο. Η διαδικασία αυτή είναι

απαραίτητη για τη σύνοψη και την κατανόηση μεγάλου όγκου δεδομένων κειμένου, επιτρέποντας την αποτελεσματική ανάκτηση και ανάλυση πληροφοριών. Στο πλαίσιο της ανάλυσης της αγοράς εργασίας, η εξαγωγή λέξεων-κλειδιών διαδραματίζει κρίσιμο ρόλο στον εντοπισμό βασικών δεξιοτήτων, προσόντων και άλλων σχετικών όρων από περιγραφές θέσεων εργασίας και βιογραφικά σημειώματα (Sali & Erden, 2022). Η ικανότητα αυτή ενισχύει την ικανότητα των εργοδοτών και των ατόμων που αναζητούν εργασία να αντιστοιχίζουν αποτελεσματικότερα τις απαιτήσεις της εργασίας με τα προφίλ των υποψηφίων (Liu et al., 2020).

Η διαδικασία εξαγωγής λέξεων-κλειδιών αρχίζει με την προεπεξεργασία των δεδομένων κειμένου. Αυτό το βήμα περιλαμβάνει τον καθαρισμό του κειμένου για την αφαίρεση άσχετων πληροφοριών, όπως τα σημεία στίξης και οι λέξεις στάσης (κοινές λέξεις όπως "και", "το" και "του" που δεν έχουν σημαντικό νόημα). Η προεπεξεργασία μπορεί επίσης να περιλαμβάνει tokenization, η οποία χωρίζει το κείμενο σε μεμονωμένες λέξεις ή φράσεις, και stemming ή lemmatization, η οποία μειώνει τις λέξεις στη ρίζα τους. Αυτά τα βήματα είναι απαραίτητα για να διασφαλιστεί ότι το κείμενο βρίσκεται σε μορφή κατάλληλη για περαιτέρω ανάλυση (Khan et al., 2022).

Αφού γίνει προεπεξεργασία του κειμένου, μπορούν να εφαρμοστούν διάφορες τεχνικές για τον εντοπισμό σημαντικών λέξεων και φράσεων. Μια συνήθης προσέγγιση είναι η χρήση στατιστικών μεθόδων, όπως η συχνότητα όρων-αντίστροφης συχνότητας εγγράφων (TF-IDF). Η TF-IDF μετρά τη σημασία μιας λέξης σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων (corpus). Η συνιστώσα συχνότητας όρων (TF) μετράει πόσο συχνά εμφανίζεται μια λέξη σε ένα έγγραφο, ενώ η συνιστώσα αντίστροφης συχνότητας εγγράφων (IDF) μειώνει τη βαρύτητα των κοινών λέξεων που εμφανίζονται σε πολλά έγγραφα. Συνδυάζοντας αυτά τα δύο μέτρα, το TF-IDF αναδεικνύει λέξεις που είναι τόσο συχνές σε ένα συγκεκριμένο έγγραφο όσο και μοναδικές σε όλο το σώμα, καθιστώντας τις πολύτιμες λέξεις-κλειδιά (Lahiri, 2019).

Για παράδειγμα, σε μια περιγραφή εργασίας για μια θέση προγραμματιστή λογισμικού, λέξεις όπως "Java", "Python", "ανάπτυξη λογισμικού" και "ευέλικτες μεθοδολογίες" μπορεί να έχουν υψηλή βαθμολογία TF-IDF. Αυτές οι λέξεις-κλειδιά παρέχουν μια συνοπτική περίληψη των βασικών δεξιοτήτων και προσόντων που απαιτούνται για τη θέση εργασίας, διευκολύνοντας τη γρήγορη και ακριβή αντιστοίχιση με προφίλ υποψηφίων που περιέχουν παρόμοιους όρους (Sali & Erden, 2022).

Μια άλλη προσέγγιση για την εξαγωγή λέξεων-κλειδιών περιλαμβάνει τη χρήση αλγορίθμων μηχανικής μάθησης. Αυτοί οι αλγόριθμοι μπορούν να εκπαιδευτούν σε σύνολα δεδομένων με ετικέτες, όπου σημαντικές λέξεις-κλειδιά σε δείγματα

κειμένων έχουν εντοπιστεί χειροκίνητα. Μαθαίνοντας από αυτά τα παραδείγματα, τα μοντέλα μηχανικής μάθησης μπορούν να προβλέψουν και να εξαγάγουν λέξεις-κλειδιά από νέα, αθέατα κείμενα. Για το σκοπό αυτό έχουν χρησιμοποιηθεί αλγόριθμοι όπως οι μηχανές διανυσμάτων υποστήριξης (SVM), τα τυχαία δάση και τα νευρωνικά δίκτυα. Αυτά τα μοντέλα μπορούν να συλλάβουν πιο σύνθετα μοτίβα και σχέσεις στο κείμενο, οδηγώντας σε ακριβέστερη εξαγωγή λέξεων-κλειδιών (Hu et al., 2022).

Τα τελευταία χρόνια, τα μοντέλα βαθιάς μάθησης, ιδίως τα μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT και το GPT, έχουν φέρει επανάσταση στην εξαγωγή λέξεων-κλειδιών. Αυτά τα μοντέλα έχουν προεκπαιδευτεί σε τεράστιες ποσότητες δεδομένων κειμένου και μπορούν να συλλάβουν βαθιές σχέσεις πλαισίου μεταξύ των λέξεων. Με τη λεπτομερή ρύθμιση αυτών των μοντέλων σε συγκεκριμένα σύνολα δεδομένων, μπορούν να χρησιμοποιηθούν για την εξαγωγή λέξεων-κλειδιών με υψηλή ακρίβεια και συνάφεια. Για παράδειγμα, το BERT μπορεί να εντοπίσει όχι μόνο μεμονωμένες λέξεις-κλειδιά αλλά και φράσεις με πολλές λέξεις και όρους με σημασία για το περιεχόμενο που μπορεί να αγνοηθούν από απλούστερες μεθόδους (Lothritz et al., 2020).

Η εφαρμογή της εξαγωγής λέξεων-κλειδιών στην ανάλυση της αγοράς εργασίας είναι ιδιαίτερα σημαντική. Εξάγοντας βασικές δεξιότητες και προσόντα από τις περιγραφές θέσεων εργασίας, οι εργοδότες μπορούν να δημιουργήσουν πιο λεπτομερείς και ακριβείς προκηρύξεις θέσεων εργασίας. Αυτή η σαφήνεια συμβάλλει στην προσέλκυση υποψηφίων που διαθέτουν τις συγκεκριμένες δεξιότητες που απαιτούνται για τη θέση. Επιπλέον, η εξαγωγή λέξεων-κλειδιών μπορεί να χρησιμοποιηθεί για την ανάλυση μεγάλων συνόλων δεδομένων περιγραφών θέσεων εργασίας για τον εντοπισμό αναδυόμενων τάσεων στη ζήτηση δεξιοτήτων. Για παράδειγμα, η αυξανόμενη συχνότητα λέξεων-κλειδιών όπως "μηχανική μάθηση", "επιστήμη των δεδομένων" και "υπολογιστικό νέφος" στις προκηρύξεις θέσεων εργασίας μπορεί να υποδηλώνει αυξανόμενη ζήτηση για αυτές τις δεξιότητες στην αγορά εργασίας (Lahiri, 2019).

Ομοίως, η εξαγωγή λέξεων-κλειδιών από βιογραφικά σημειώματα επιτρέπει τη δημιουργία δομημένων προφίλ υποψηφίων. Με τον εντοπισμό των βασικών δεξιοτήτων και προσόντων στα βιογραφικά σημειώματα, οι υπεύθυνοι προσλήψεων μπορούν να αντιστοιχίσουν γρήγορα τους υποψηφίους με τις σχετικές θέσεις εργασίας. Αυτή η αυτοματοποίηση μειώνει το χρόνο και την προσπάθεια που απαιτούνται για τη χειροκίνητη εξέταση βιογραφικών σημειωμάτων και βελτιώνει την ακρίβεια της αντιστοίχισης υποψηφίων και θέσεων εργασίας. Για παράδειγμα, ένα βιογραφικό σημείωμα που περιέχει λέξεις-κλειδιά όπως "διαχείριση έργου", "προγραμματισμός προϋπολογισμού" και

"επικοινωνία με τους ενδιαφερόμενους" μπορεί εύκολα να αντιστοιχιστεί με μια περιγραφή θέσης εργασίας για έναν ρόλο διαχειριστή έργου (Khan et al., 2022).

Τεχνικές και μοντέλα

Η εξαγωγή λέξεων-κλειδιών είναι μια κρίσιμη εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που επικεντρώνεται στον εντοπισμό σημαντικών λέξεων και φράσεων μέσα σε ένα κείμενο. Η διαδικασία αυτή είναι απαραίτητη για τη σύνοψη και την κατηγοριοποίηση μεγάλου όγκου δεδομένων κειμένου, η οποία είναι ιδιαίτερα χρήσιμη στην ανάλυση της αγοράς εργασίας, όπου η κατανόηση των περιγραφών θέσεων εργασίας και των βιογραφικών σημειωμάτων είναι ζωτικής σημασίας. Έχουν αναπτυχθεί διάφορες τεχνικές και μοντέλα για την ενίσχυση της ακρίβειας και της αποτελεσματικότητας της εξαγωγής λέξεων-κλειδιών, συμπεριλαμβανομένων στατιστικών μεθόδων, προσεγγίσεων μηχανικής μάθησης και προηγμένων αλγορίθμων (Lothritz et al., 2020).

Οι στατιστικές μέθοδοι είναι από τις πρώτες τεχνικές που χρησιμοποιήθηκαν για την εξαγωγή λέξεων-κλειδιών. Μια από τις πιο γνωστές στατιστικές μεθόδους είναι η μέθοδος Term Frequency-Inverse Document Frequency (TF-IDF). Η TF-IDF αξιολογεί τη σημασία μιας λέξης σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων (Ganiger & Rajashekharaiyah, 2018). Η συνιστώσα συχνότητας όρων μετρά πόσο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο, ενώ η συνιστώσα αντίστροφης συχνότητας εγγράφων μειώνει τη βαρύτητα των όρων που εμφανίζονται συχνά σε πολλά έγγραφα. Αυτό το συνδυασμένο μέτρο βοηθά στην ανάδειξη όρων που είναι ιδιαίτερα σημαντικοί για ένα συγκεκριμένο έγγραφο. Για παράδειγμα, σε μια περιγραφή θέσης εργασίας για έναν επιστήμονα δεδομένων, όροι όπως "Python", "μηχανική μάθηση" και "ανάλυση δεδομένων" μπορεί να λάβουν υψηλή βαθμολογία TF-IDF, υποδεικνύοντας τη σημασία τους για το έγγραφο (Campros et al., 2018).

Μια άλλη αποτελεσματική στατιστική μέθοδος είναι η ταχεία αυτόματη εξαγωγή λέξεων-κλειδιών (RAKE). Η RAKE λειτουργεί με τον εντοπισμό υποψηφίων λέξεων-κλειδιών σε ένα κείμενο και στη συνέχεια με την ανάλυση της συχνότητας και της συνύπαρξης αυτών των υποψηφίων λέξεων-κλειδιών για τον προσδιορισμό της σημασίας τους. Σε αντίθεση με την TF-IDF, η οποία βασίζεται σε ένα σώμα εγγράφων, η RAKE λειτουργεί σε μεμονωμένα έγγραφα, γεγονός που την καθιστά χρήσιμη για την εξαγωγή λέξεων-κλειδιών από μεμονωμένες περιγραφές θέσεων εργασίας ή βιογραφικά σημειώματα. Το RAKE τμηματοποιεί το κείμενο σε φράσεις με βάση οριοθέτες, όπως σημεία στίξης και λέξεις παύσης, και στη συνέχεια υπολογίζει μια βαθμολογία για κάθε φράση με βάση τη συχνότητα των λέξεων που τη συνθέτουν και τη συνύπαρξή τους. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη για τον εντοπισμό πολυλεκτικών εκφράσεων, οι οποίες είναι συχνές σε λεπτομερείς περιγραφές θέσεων εργασίας (Cai & Cao, 2017).

Οι προσεγγίσεις μηχανικής μάθησης για την εξαγωγή λέξεων-κλειδιών έχουν κερδίσει δημοτικότητα λόγω της ικανότητάς τους να μαθαίνουν από τα δεδομένα και να βελτιώνονται με την πάροδο του χρόνου. Αυτές οι προσεγγίσεις περιλαμβάνουν την εκπαίδευση μοντέλων σε σχολιασμένα σύνολα δεδομένων όπου οι λέξεις-κλειδιά έχουν εντοπιστεί χειροκίνητα (Thiyagarajan et al., 2021). Ένας συνηθισμένος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για τον σκοπό αυτό είναι ο TextRank, ένα μοντέλο κατάταξης βασισμένο σε γράφους εμπνευσμένο από τον αλγόριθμο PageRank που χρησιμοποιείται από την Google. Ο TextRank κατασκευάζει έναν γράφο όπου οι κόμβοι αντιπροσωπεύουν λέξεις και οι ακμές αντιπροσωπεύουν συν-εμφανίσεις εντός ενός συγκεκριμένου παραθύρου λέξεων. Εφαρμόζοντας τον αλγόριθμο PageRank σε αυτόν τον γράφο, ο TextRank μπορεί να κατατάξει τις λέξεις με βάση τη σημασία τους μέσα στο κείμενο, εξάγοντας τις λέξεις με την υψηλότερη κατάταξη ως λέξεις-κλειδιά (Campos et al., 2018). Το TextRank είναι ιδιαίτερα αποτελεσματικό για την καταγραφή των σχέσεων μεταξύ των λέξεων, καθιστώντας το μια ισχυρή μέθοδο για την εξαγωγή λέξεων-κλειδιών (Hu et al., 2022).

Τα μοντέλα βαθιάς μάθησης, ειδικά τα μοντέλα που βασίζονται σε μετασχηματιστές όπως το BERT (Bidirectional Encoder Representations from Transformers), αντιπροσωπεύουν την τελευταία εξέλιξη στην εξαγωγή λέξεων-κλειδιών. Το BERT προ-εκπαιδεύεται σε ένα τεράστιο σώμα κειμένων και τελειοποιείται σε συγκεκριμένες εργασίες, επιτρέποντάς του να συλλαμβάνει βαθιές πληροφορίες πλαισίου και σχέσεις μεταξύ των λέξεων. Για την εξαγωγή λέξεων-κλειδιών, η BERT μπορεί να ρυθμιστεί λεπτομερώς σε σύνολα δεδομένων με σχόλια για τον εντοπισμό λέξεων-κλειδιών με υψηλή ακρίβεια (Ganiger & Rajashekharaiyah, 2018). Η αμφίδρομη φύση του BERT του επιτρέπει να εξετάζει το πλαίσιο και από τις δύο κατευθύνσεις σε μια ακολουθία κειμένου, καθιστώντας το ιδιαίτερα αποτελεσματικό στην κατανόηση των αποχρώσεων της γλώσσας. Για παράδειγμα, η BERT μπορεί να εξάγει με ακρίβεια πολύπλοκες δεξιότητες και προσόντα από περιγραφές θέσεων εργασίας, όπως "εμπειρία με πλαίσια βαθιάς μάθησης" ή "επάρκεια στη στατιστική μοντελοποίηση" (Lothritz et al., 2020).

1.4. Λεπτομερής συζήτηση σχετικά με την NER στην ανάλυση της αγοράς εργασίας

1.4.1. Τεχνικές NER για την εξαγωγή δεξιοτήτων

Προσεγγίσεις βασισμένες σε κανόνες

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι μια βασική εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την ταξινόμηση οντοτήτων σε ένα κείμενο. Στο πλαίσιο της ανάλυσης της αγοράς εργασίας, οι τεχνικές NER χρησιμοποιούνται για την εξαγωγή σχετικών δεξιοτήτων από βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας, βοηθώντας στην αποτελεσματική αντιστοίχιση υποψηφίων με ευκαιρίες απασχόλησης. Μια από τις θεμελιώδεις προσεγγίσεις της NER για την εξαγωγή δεξιοτήτων είναι η χρήση μεθόδων βασισμένων σε κανόνες. Αυτές οι μέθοδοι βασίζονται σε προκαθορισμένους κανόνες για τον εντοπισμό οντοτήτων μέσα στο κείμενο, προσφέροντας ένα απλό και ερμηνεύσιμο μέσο για την εξαγωγή συγκεκριμένων τύπων πληροφοριών (Liu et al., 2020).

Οι βασισμένες σε κανόνες προσεγγίσεις στην ΕΑΑ περιλαμβάνουν τη δημιουργία προτύπων και κανόνων που χρησιμοποιεί το σύστημα για τον εντοπισμό και την εξαγωγή οντοτήτων. Αυτοί οι κανόνες συχνά διατυπώνονται με βάση τα γλωσσικά χαρακτηριστικά των οντοτήτων ενδιαφέροντος (Mulalic, 2019). Για παράδειγμα, για την εξαγωγή γλωσσών προγραμματισμού από ένα βιογραφικό σημείωμα, μπορεί να δημιουργηθούν κανόνες που αναγνωρίζουν συγκεκριμένες λέξεις-κλειδιά όπως "Python", "Java", "C++" και "JavaScript". Στη συνέχεια, το σύστημα σαρώνει το κείμενο για αυτές τις λέξεις-κλειδιά, εφαρμόζοντας τους κανόνες για να τις αναγνωρίσει και να τις ταξινομήσει ως γλώσσες προγραμματισμού (Todorovic et al., 2010).

Ένα από τα πιο συνηθισμένα εργαλεία που χρησιμοποιούνται στη NER βάσει κανόνων είναι οι κανονικές εκφράσεις. Οι κανονικές εκφράσεις είναι ακολουθίες χαρακτήρων που ορίζουν μοτίβα αναζήτησης, τα οποία μπορούν να χρησιμοποιηθούν για την αντιστοίχιση και την εξαγωγή συγκεκριμένων συμβολοσειρών κειμένου. Είναι ιδιαίτερα αποτελεσματικές για τον εντοπισμό καλά καθορισμένων οντοτήτων, όπως ημερομηνίες, διευθύνσεις ηλεκτρονικού ταχυδρομείου και γλώσσες προγραμματισμού (Campos et al., 2018). Για παράδειγμα, μια κανονική έκφραση που έχει σχεδιαστεί για την εξαγωγή γλωσσών προγραμματισμού μπορεί να αναζητήσει λέξεις που εμφανίζονται σε έναν κατάλογο γνωστών γλωσσών προγραμματισμού ή να ταιριάζει με μοτίβα που είναι τυπικά για τα ονόματα γλωσσών προγραμματισμού. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για την εξαγωγή οντοτήτων από κείμενα με προβλέψιμες και δομημένες μορφές.

Τα πλεονεκτήματα των προσεγγίσεων που βασίζονται σε κανόνες έγκεινται στην απλότητα και την ακρίβειά τους. Δεδομένου ότι οι κανόνες ορίζονται ρητά, το σύστημα μπορεί να επιτύχει υψηλή ακρίβεια στην αναγνώριση οντοτήτων που ταιριάζουν στα προκαθορισμένα πρότυπα. Αυτό είναι ιδιαίτερα επωφελές σε τομείς όπου οι οντότητες ενδιαφέροντος είναι γνωστές και έχουν συγκεκριμένα,

αναγνωρίσιμα χαρακτηριστικά (Thiyagarajan et al., 2021). Για παράδειγμα, η εξαγωγή γλωσσών προγραμματισμού από ένα βιογραφικό σημείωμα είναι σχετικά απλή, επειδή τα ονόματα των γλωσσών προγραμματισμού είναι διακριτά και τυποποιημένα (Mulalic, 2019).

Ωστόσο, οι προσεγγίσεις που βασίζονται σε κανόνες έχουν επίσης αξιοσημείωτους περιορισμούς. Ένα σημαντικό μειονέκτημα είναι η έλλειψη ευελιξίας και προσαρμοστικότητας. Επειδή οι κανόνες είναι χειροποίητοι, μπορεί να δυσκολεύονται με τις παραλλαγές στη χρήση της γλώσσας, όπως συνώνυμα, ορθογραφικά λάθη και συντομογραφίες. Για παράδειγμα, αν σε ένα βιογραφικό σημείωμα αναφέρεται "JS" αντί για "JavaScript", ένα σύστημα που βασίζεται σε κανόνες και αναζητά μόνο τον πλήρη όρο μπορεί να μην εντοπίσει αυτή την περίπτωση. Επιπλέον, τα συστήματα που βασίζονται σε κανόνες μπορεί να γίνουν δυσκίνητα στη συντήρηση καθώς αυξάνεται ο αριθμός των κανόνων, ιδίως σε δυναμικά πεδία όπου εμφανίζονται συχνά νέες οντότητες (Liu et al., 2020).

Στατιστικές μέθοδοι

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι μια κρίσιμη εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την ταξινόμηση οντοτήτων σε κείμενο. Στην ανάλυση της αγοράς εργασίας, οι τεχνικές NER είναι ζωτικής σημασίας για την εξαγωγή δεξιοτήτων, τίτλων θέσεων εργασίας και άλλων σχετικών πληροφοριών από βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας. Μία από τις προηγμένες μεθόδους για την NER περιλαμβάνει τη χρήση στατιστικών προσεγγίσεων, ιδίως πιθανοτικών μοντέλων (Colmenar et al., 2011). Μεταξύ αυτών, τα κρυφά μοντέλα Markov (Hidden Markov Models - HMMs) έχουν χρησιμοποιηθεί σε περίοπτη θέση για την αναγνώριση οντοτήτων με βάση το πλαίσιο, παρέχοντας ένα ισχυρό πλαίσιο για τον εντοπισμό τίτλων εργασίας και άλλων οντοτήτων (Wang et al., 2018).

Οι στατιστικές μέθοδοι, συμπεριλαμβανομένων των HMM, λειτουργούν με βάση την αρχή της πιθανοτικής μοντελοποίησης. Οι HMM είναι ιδιαίτερα αποτελεσματικές για δεδομένα ακολουθίας, όπως το κείμενο, όπου ο στόχος είναι να επισημανθεί κάθε λέξη ή συμβολισμός σε μια ακολουθία. Μια HMM είναι ένα στατιστικό μοντέλο που υποθέτει ότι το σύστημα που μοντελοποιείται είναι μια διαδικασία Markov με κρυφές καταστάσεις. Στο πλαίσιο της NER, οι κρυφές καταστάσεις αντιστοιχούν στους διάφορους τύπους οντοτήτων (όπως τίτλοι θέσεων εργασίας, δεξιότητες ή οργανισμοί), ενώ οι παρατηρούμενες καταστάσεις είναι οι λέξεις του κειμένου (Zafarian et al., 2015).

Η διαδικασία χρήσης των HMMs για NER περιλαμβάνει διάφορα βήματα. Πρώτον, το μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων με ετικέτες, όπου κάθε λέξη του κειμένου επισημαίνεται με τον αντίστοιχο τύπο οντότητας. Κατά τη διάρκεια

της εκπαίδευσης, το μοντέλο μαθαίνει τις πιθανότητες των μεταβάσεων μεταξύ καταστάσεων (δηλαδή την πιθανότητα ένας συγκεκριμένος τύπος οντότητας να ακολουθεί έναν άλλο) και τις πιθανότητες εκπομπής κάθε λέξης δεδομένης μιας συγκεκριμένης κατάστασης (δηλαδή την πιθανότητα μια λέξη να συσχετίζεται με έναν συγκεκριμένο τύπο οντότητας). Αυτή η διαδικασία εκπαίδευσης περιλαμβάνει την εκτίμηση δύο συνόλων πιθανοτήτων: των πιθανοτήτων μετάβασης και των πιθανοτήτων εκπομπής (Wang et al., 2018).

Αφού εκπαιδευτεί, το HMM μπορεί να εφαρμοστεί σε νέο, άρατο κείμενο για τον εντοπισμό οντοτήτων. Δεδομένης μιας ακολουθίας λέξεων (όπως μια πρόταση από ένα βιογραφικό σημείωμα), το μοντέλο χρησιμοποιεί τις πιθανότητες που έχει μάθει για να συμπεράνει την πιο πιθανή ακολουθία κρυφών καταστάσεων (τύποι οντοτήτων). Αυτή η εξαγωγή συμπερασμάτων γίνεται συνήθως με τη χρήση του αλγορίθμου Viterbi, ο οποίος βρίσκει αποτελεσματικά την πιο πιθανή ακολουθία καταστάσεων (Colmenar et al., 2011).

Προσεγγίσεις μηχανικής μάθησης

Η αναγνώριση ονομαστικών οντοτήτων (NER) έχει εξελιχθεί σημαντικά με την έλευση προσεγγίσεων μηχανικής μάθησης, οι οποίες προσφέρουν πιο ισχυρές και προσαρμόσιμες μεθόδους για την εξαγωγή οντοτήτων από κείμενο. Στο πλαίσιο της ανάλυσης της αγοράς εργασίας, οι τεχνικές NER που βασίζονται στη μηχανική μάθηση είναι ιδιαίτερα αποτελεσματικές για την εξαγωγή δεξιοτήτων από βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας. Μεταξύ αυτών των τεχνικών, οι μέθοδοι μάθησης με επίβλεψη έχουν αποδειχθεί ιδιαίτερα επιτυχείς. Αυτές οι μέθοδοι περιλαμβάνουν την εκπαίδευση μοντέλων σε σχολιασμένα σύνολα δεδομένων όπου οι οντότητες έχουν επισημανθεί χειροκίνητα. Δύο εξέχοντα παραδείγματα τέτοιων προσεγγίσεων είναι τα Conditional Random Fields (CRF) και τα δίκτυα αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (BiLSTM) (Li et al., 2021).

Τα Conditional Random Fields (CRF) είναι ένας τύπος πιθανοτικού γραφικού μοντέλου που χρησιμοποιείται για δομημένη πρόβλεψη. Στο πλαίσιο της NER, τα CRF χρησιμοποιούνται για την επισήμανση ακολουθιών λέξεων, λαμβάνοντας υπόψη το πλαίσιο που παρέχουν οι γειτονικές λέξεις. Η ισχύς των CRF έγκειται στην ικανότητά τους να μοντελοποιούν τις εξαρτήσεις μεταξύ των ετικετών, η οποία είναι ζωτικής σημασίας για τον ακριβή εντοπισμό οντοτήτων στο κείμενο. Όταν εφαρμόζεται στην εξαγωγή δεξιοτήτων, ένα μοντέλο CRF μπορεί να εκπαιδευτεί σε σχολιασμένα βιογραφικά σημειώματα όπου οι δεξιότητες επισημαίνονται ρητά. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει τις πιθανότητες διαφορετικών ακολουθιών ετικετών, επιτρέποντάς του να προβλέψει την πιο πιθανή ακολουθία ετικετών για νέα, αθέατα βιογραφικά σημειώματα (Gao et al., 2021). Για παράδειγμα, ένα CRF μπορεί να εκπαιδευτεί

για να αναγνωρίζει τις δεξιότητες "Python", "διαχείριση έργων" και "ανάλυση δεδομένων", αξιοποιώντας το πλαίσιο στο οποίο εμφανίζονται αυτοί οι όροι για να αυξήσει την ακρίβεια (Sun et al., 2017).

Τα δίκτυα αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (BiLSTM) αντιπροσωπεύουν μια πιο προηγμένη προσέγγιση μηχανικής μάθησης που έχει συγκεντρώσει σημαντική προσοχή τα τελευταία χρόνια. Τα BiLSTM είναι ένας τύπος επαναλαμβανόμενου νευρωνικού δικτύου (RNN) που επεξεργάζεται ακολουθίες κειμένου τόσο προς τα εμπρός όσο και προς τα πίσω. Αυτή η αμφίδρομη ικανότητα επιτρέπει στα BiLSTMs να συλλαμβάνουν τα συμφραζόμενα τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις, παρέχοντας μια ολοκληρωμένη κατανόηση του κειμένου. Όταν χρησιμοποιούνται για NER, τα BiLSTM είναι ιδιαίτερα ικανά στην αναγνώριση οντοτήτων που εξαρτώνται από το περιεχόμενο (Mulalic, 2019). Για παράδειγμα, ο όρος "Java" μπορεί να αναφέρεται σε μια γλώσσα προγραμματισμού ή σε ένα νησί, ανάλογα με τις λέξεις που τον περιβάλλουν. Επεξεργαζόμενοι το κείμενο αμφίδρομα, οι BiLSTM μπορούν να ξεχωρίσουν με ακρίβεια τέτοιους όρους με βάση τα συμφραζόμενα.

Η διαδικασία εκπαίδευσης για τα BiLSTM περιλαμβάνει την εισαγωγή σχολιασμένου κειμένου στο μοντέλο, όπου κάθε λέξη επισημαίνεται με τον αντίστοιχο τύπο οντότητας. Το δίκτυο μαθαίνει να προβλέπει αυτές τις ετικέτες ρυθμίζοντας τις εσωτερικές του παραμέτρους μέσω οπισθοδιάδοσης. Αφού εκπαιδευτεί, το BiLSTM μπορεί να εφαρμοστεί σε νέο κείμενο για την εξαγωγή οντοτήτων. Στην περίπτωση της εξαγωγής δεξιοτήτων από βιογραφικά σημειώματα, ένα BiLSTM μπορεί να εντοπίσει σύνθετες και πολυλεκτικές δεξιότητες, όπως "αλγόριθμοι μηχανικής μάθησης" ή "αξιολόγηση χρηματοοικονομικών κινδύνων", κατανοώντας το ευρύτερο πλαίσιο στο οποίο αναφέρονται αυτές οι δεξιότητες (Wang et al., 2018).

Ένα από τα σημαντικά πλεονεκτήματα της χρήσης μεθόδων μάθησης με επίβλεψη, όπως η CRF και η BiLSTM, για την εξαγωγή δεξιοτήτων είναι η ικανότητά τους να γενικεύουν από σχολιασμένα δεδομένα εκπαίδευσης σε νέο, αθέατο κείμενο. Αυτή η ικανότητα είναι ιδιαίτερα χρήσιμη στην αγορά εργασίας, όπου η γλώσσα που χρησιμοποιείται στα βιογραφικά σημειώματα και στις περιγραφές θέσεων εργασίας μπορεί να ποικίλλει σε μεγάλο βαθμό (Sun et al., 2017). Με την εκπαίδευση σε διαφορετικά σύνολα δεδομένων, τα μοντέλα αυτά μπορούν να μάθουν να αναγνωρίζουν ένα ευρύ φάσμα δεξιοτήτων και να προσαρμόζονται σε διαφορετικά στυλ γραφής και ορολογίες. Αυτή η προσαρμοστικότητα διασφαλίζει ότι τα μοντέλα παραμένουν αποτελεσματικά ακόμη και όταν η αγορά εργασίας εξελίσσεται και εμφανίζονται νέες δεξιότητες (Li et al., 2021).

Επιπλέον, η χρήση της μάθησης με επίβλεψη για τη NER διευκολύνει τη συνεχή βελτίωση και τελειοποίηση. Καθώς διατίθενται περισσότερα σχολιασμένα δεδομένα, τα μοντέλα μπορούν να επανεκπαιδευτούν για να βελτιώσουν την ακρίβεια και την κάλυψή τους. Αυτή η επαναληπτική διαδικασία επιτρέπει στους οργανισμούς να διατηρούν τα συστήματά τους NER ενημερωμένα με τις τελευταίες τάσεις στη ζήτηση δεξιοτήτων και τις απαιτήσεις της αγοράς εργασίας (Gao et al., 2021).

Βαθιά μάθηση και μετασχηματιστές

Η αναγνώριση ονομαστικών οντοτήτων (NER) έχει ενισχυθεί σημαντικά με την έλευση της βαθιάς μάθησης και των μοντέλων μετασχηματισμού, τα οποία προσφέρουν εξελιγμένες μεθόδους για τη σύλληψη του πλαισίου και των εξαρτήσεων στο κείμενο. Αυτά τα προηγμένα μοντέλα, ιδίως το BERT (Bidirectional Encoder Representations from Transformers) και το GPT (Generative Pre-trained Transformer), έχουν θέσει νέα πρότυπα για την υψηλής ακρίβειας NER. Η ικανότητά τους να κατανοούν και να επεξεργάζονται σύνθετα γλωσσικά πρότυπα τα καθιστά ανεκτίμητα για την εξαγωγή δεξιοτήτων στην ανάλυση της αγοράς εργασίας, όπου η ακρίβεια και το πλαίσιο είναι ζωτικής σημασίας (Liu et al., 2020).

Το BERT, που αναπτύχθηκε από την Google, αποτελεί ένα σημαντικό άλμα προόδου στον τομέα του NLP. Σε αντίθεση με τα παραδοσιακά μοντέλα που επεξεργάζονται το κείμενο διαδοχικά, το BERT είναι αμφίδρομο, δηλαδή εξετάζει τα συμφραζόμενα τόσο από τα αριστερά όσο και από τα δεξιά κάθε λέξης. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο BERT να κατανοεί τις αποχρώσεις των σχέσεων μεταξύ των λέξεων και του περιβάλλοντος πλαισίου τους. Στο πλαίσιο της NER για την εξαγωγή δεξιοτήτων, αυτή η ικανότητα είναι ιδιαίτερα επωφελής. Για παράδειγμα, η BERT μπορεί να διακρίνει μεταξύ διαφορετικών χρήσεων της λέξης "Python" κατανοώντας αν περιβάλλεται από όρους που σχετίζονται με τον προγραμματισμό ή τη βιολογία. Αυτό το επίπεδο κατανόησης του πλαισίου επιτρέπει στην BERT να εντοπίζει με ακρίβεια γλώσσες προγραμματισμού, τεχνικές δεξιότητες και άλλες οντότητες σε βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας (Zhang et al., 2019).

Η διαδικασία χρήσης του BERT για NER περιλαμβάνει την προ-εκπαίδευση του μοντέλου σε ένα μεγάλο σώμα κειμένου για την εκμάθηση γενικών γλωσσικών αναπαραστάσεων, ακολουθούμενη από την εκτέλεση του σε ένα συγκεκριμένο σύνολο δεδομένων NER, όπου οι οντότητες είναι σχολιασμένες. Κατά τη διάρκεια της τελειοποίησης, το BERT μαθαίνει να αναγνωρίζει και να ταξινομεί οντότητες με βάση το πλαίσιο που παρέχει το περιβάλλον κείμενο. Αυτή η προσέγγιση εκπαίδευσης επιτρέπει στην BERT να γενικεύει καλά σε νέο, αόρατο κείμενο,

καθιστώντας την εξαιρετικά αποτελεσματική για την εξαγωγή δεξιοτήτων από ποικίλα και μεταβλητά βιογραφικά σημειώματα (Syed & Chung, 2021).

Τα μοντέλα GPT, ιδίως το GPT-3 που αναπτύχθηκε από την OpenAI, έχουν επίσης δείξει αξιοσημείωτες ικανότητες στη NER. Το GPT-3 είναι ένα μοντέλο βασισμένο σε μετασχηματιστές που χρησιμοποιεί μια μονόδρομη προσέγγιση, αλλά αντισταθμίζει με την τεράστια κλίμακα του, καθώς εκπαιδεύεται σε εκατοντάδες δισεκατομμύρια παραμέτρων. Ενώ το GPT-3 χρησιμοποιείται γενικά για τη δημιουργία κειμένων, η βαθιά κατανόηση του πλαισίου του το καθιστά κατάλληλο και για εργασίες NER. Για την εξαγωγή δεξιοτήτων, το GPT-3 μπορεί να αναλύσει περιγραφές θέσεων εργασίας και βιογραφικά σημειώματα για τον εντοπισμό σχετικών δεξιοτήτων, αξιοποιώντας την εκτεταμένη εκπαίδευσή του για να χειρίζεται ένα ευρύ φάσμα γλωσσικών παραλλαγών και αποχρώσεων (Zhang et al., 2019).

Η εφαρμογή της BERT και της GPT για NER στην ανάλυση της αγοράς εργασίας προσφέρει πολλά βασικά πλεονεκτήματα. Ένα σημαντικό πλεονέκτημα είναι η ικανότητά τους να χειρίζονται πολύπλοκα και διαφορούμενα κείμενα. Τα βιογραφικά σημειώματα συχνά περιέχουν ποικίλη και μη τυποποιημένη γλώσσα για την περιγραφή παρόμοιων δεξιοτήτων και εμπειριών (Syed & Chung, 2021). Τα παραδοσιακά μοντέλα NER μπορεί να δυσκολεύονται με αυτή τη μεταβλητότητα, αλλά τα BERT και GPT μπορούν να ερμηνεύσουν με ακρίβεια και να εξάγουν τις απαραίτητες πληροφορίες λόγω της βαθιάς τους επίγνωσης του πλαισίου. Για παράδειγμα, φράσεις όπως "ικανός στην ανάλυση δεδομένων με χρήση Python" και "έμπειρος με Python για την επιστήμη των δεδομένων" μπορεί να διαφέρουν στη διατύπωση αλλά να μεταφέρουν την ίδια δεξιότητα. Οι BERT και GPT μπορούν να αναγνωρίσουν αυτή την ισοδυναμία και να ταξινομήσουν σωστά και τις δύο φράσεις κάτω από τη δεξιότητα "προγραμματισμός Python" (Liu et al., 2020).

1.4.2. Προκλήσεις στην NER για την ανάλυση βιογραφικού σημειώματος

Ποικίλες μορφές

Η αναγνώριση ονομαστικών οντοτήτων (NER) για την ανάλυση βιογραφικών σημειωμάτων διαδραματίζει κρίσιμο ρόλο στην αυτοματοποιημένη εξαγωγή σχετικών πληροφοριών από βιογραφικά σημειώματα, βοηθώντας σε αποτελεσματικές διαδικασίες πρόσληψης. Ωστόσο, μία από τις σημαντικές προκλήσεις που αντιμετωπίζουν τα συστήματα NER σε αυτό το πλαίσιο είναι ο χειρισμός των ποικίλων μορφών των βιογραφικών σημειωμάτων. Τα βιογραφικά σημειώματα μπορεί να έχουν πολλαπλές δομές και διατάξεις, όπως PDF, έγγραφα

του Word και μορφές απλού κειμένου, που το καθένα παρουσιάζει μοναδικά εμπόδια για τους αλγορίθμους NER (Sougandh et al., 2023).

Τα βιογραφικά σημειώματα σε μορφή PDF αποτελούν ιδιαίτερη πρόκληση λόγω του τρόπου κωδικοποίησης του κειμένου στα PDF. Σε αντίθεση με το απλό κείμενο, όπου το περιεχόμενο αποθηκεύεται διαδοχικά, τα PDF μπορούν να έχουν κείμενο τοποθετημένο σε πολύπλοκες διατάξεις, συμπεριλαμβανομένων στηλών, πινάκων και ενσωματωμένων εικόνων. Η διαδικασία εξαγωγής κειμένου από αρχεία PDF περιλαμβάνει συχνά τη μετατροπή του εγγράφου σε μορφή αναγνώσιμη από μηχανήματα, η οποία μπορεί να οδηγήσει σε σφάλματα όπως εσφαλμένη ευθυγράμμιση του κειμένου, απώλεια μορφοποίησης και έλλειψη δεδομένων. Αυτά τα ζητήματα περιπλέκουν τη διαδικασία NER, καθώς το μοντέλο μπορεί να δυσκολεύεται να εντοπίσει και να ταξινομήσει σωστά οντότητες όπως τίτλους θέσεων εργασίας, ονόματα εταιρειών και δεξιότητες μέσα σε μια ακατάστατη εξαγωγή κειμένου (Tiedemann, 2014).

Για παράδειγμα, σκεφτείτε ένα βιογραφικό σημείωμα με διάταξη δύο στηλών, όπου τα προσωπικά στοιχεία βρίσκονται στα αριστερά και η επαγγελματική εμπειρία στα δεξιά. Κατά τη διαδικασία μετατροπής, το κείμενο και από τις δύο στήλες ενδέχεται να συγχωνευτεί εσφαλμένα, οδηγώντας σε μη λογικές ακολουθίες που εμποδίζουν την αναγνώριση οντοτήτων. Για την ακριβή μετατροπή περιεχομένου PDF σε δομημένη μορφή κατάλληλη για NER απαιτούνται συχνά προηγμένα βήματα προεπεξεργασίας, όπως ανάλυση διάταξης και οπτική αναγνώριση χαρακτήρων (OCR). Παρά τις προσπάθειες αυτές, η ασυνέπεια στις δομές PDF μπορεί να εξακολουθεί να αποτελεί σημαντική πρόκληση για την αξιόπιστη εξαγωγή οντοτήτων (Anantharaman et al., 2022).

Τα έγγραφα του Word, αν και γενικά είναι πιο απλά από τα PDF, παρουσιάζουν επίσης προκλήσεις λόγω της ποικίλης χρήσης χαρακτηριστικών μορφοποίησης, όπως πίνακες, κουκκίδες και διαφορετικά στυλ γραμματοσειράς. Κάθε βιογραφικό σημείωμα μπορεί να χρησιμοποιεί έναν μοναδικό συνδυασμό αυτών των χαρακτηριστικών, γεγονός που καθιστά δύσκολο για τα συστήματα NER να αναγνωρίζουν με συνέπεια τις οντότητες. Για παράδειγμα, ένα βιογραφικό σημείωμα μπορεί να απαριθμεί τις δεξιότητες σε μορφή πίνακα, με γραμμές και στήλες που πρέπει να ερμηνευθούν σωστά για να εξαχθούν οι σχετικές πληροφορίες. Επιπλέον, τα έγγραφα του Word μπορεί να περιέχουν παρακολουθούμενες αλλαγές, σχόλια και άλλα μεταδεδομένα που πρέπει να φιλτράρονται κατά την προεπεξεργασία για να μην μπερδεύεται το μοντέλο NER (Tiedemann, 2014).

Ο χειρισμός βιογραφικών σημειωμάτων απλού κειμένου μπορεί να φαίνεται απλούστερος με την πρώτη ματιά, αλλά ακόμη και αυτά μπορεί να διαφέρουν σημαντικά ως προς τη δομή τους. Τα βιογραφικά σημειώματα απλού κειμένου

δεν διαθέτουν τις πλούσιες επιλογές μορφοποίησης των PDF και των εγγράφων του Word, με αποτέλεσμα οι υποψήφιοι να χρησιμοποιούν διάφορους διαχωριστές, στυλ εσοχών και διαλείμματα γραμμών για να οργανώσουν τις πληροφορίες τους. Η απουσία τυποποιημένης μορφοποίησης σημαίνει ότι τα συστήματα NER πρέπει να βασίζονται σε μεγάλο βαθμό στο πλαίσιο και στην αναγνώριση προτύπων για τον σωστό εντοπισμό οντοτήτων. Για παράδειγμα, η διάκριση μεταξύ τμημάτων όπως "Εμπειρία" και "Εκπαίδευση" μπορεί να αποτελέσει πρόκληση εάν δεν είναι σαφώς σημειωμένα ή εάν το βιογραφικό σημείωμα χρησιμοποιεί μη συμβατικές επικεφαλίδες (Sougandh et al., 2023).

Για την αντιμετώπιση αυτών των προκλήσεων, τα συστήματα NER συχνά ενσωματώνουν αγωγούς προεπεξεργασίας προσαρμοσμένους στις διάφορες μορφές βιογραφικών σημειωμάτων. Για τα PDF, αυτό μπορεί να περιλαμβάνει τη χρήση προηγμένων τεχνικών OCR και ανάλυσης διάταξης για την ακριβή ανακατασκευή της δομής του εγγράφου. Για έγγραφα του Word, τα βήματα προεπεξεργασίας μπορεί να περιλαμβάνουν την εξαγωγή και την κανονικοποίηση του κειμένου, ενώ παράλληλα φιλτράρονται τα άσχετα μεταδεδομένα. Τα βιογραφικά σημειώματα απλού κειμένου απαιτούν ισχυρή αντιστοίχιση μοτίβων και ανάλυση πλαισίου για τη σωστή ερμηνεία του περιεχομένου (Anantharaman et al., 2022).

Ασάφειες και συνώνυμα

Η αναγνώριση ονομαστικών οντοτήτων (NER) για την ανάλυση βιογραφικών σημειωμάτων είναι μια κρίσιμη εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP), απαραίτητη για την ακριβή εξαγωγή σχετικών πληροφοριών, όπως δεξιότητες, τίτλοι εργασίας και προσόντα. Ωστόσο, μία από τις σημαντικές προκλήσεις στην ανάλυση βιογραφικού σημειώματος είναι η αντιμετώπιση των ασάφειες και των συνωνύμων. Η εγγενής μεταβλητότητα στον τρόπο με τον οποίο τα άτομα περιγράφουν τις δεξιότητες και τις εμπειρίες τους μπορεί να περιπλέξει το έργο της συνεπούς αναγνώρισης και ταξινόμησης αυτών των οντοτήτων. Το ζήτημα αυτό είναι ιδιαίτερα εμφανές όταν χρησιμοποιούνται διαφορετικοί όροι για να αναφερθούν στην ίδια δεξιότητα, όπως "προγραμματισμός Java" και "ανάπτυξη Java" (Zhao et al., 2015).

Οι ασάφειες στα βιογραφικά προκύπτουν από τους διαφορετικούς τρόπους με τους οποίους οι υποψήφιοι περιγράφουν τις ικανότητες και τις εμπειρίες τους. Για παράδειγμα, ένας υποψήφιος μπορεί να αναφέρεται στην εξειδίκευσή του στη Java με διάφορους τρόπους, όπως "προγραμματισμός Java", "ανάπτυξη Java", "ικανός στη Java" ή "εμπειρία με Java". Παρόλο που αυτές οι φράσεις αποδίδουν την ίδια βασική δεξιότητα, οι διαφορές στη διατύπωση μπορεί να δημιουργήσουν προκλήσεις για τα συστήματα NER, τα οποία πρέπει να αναγνωρίζουν με ακρίβεια και να ταξινομούν αυτές τις παραλλαγές ως ενιαία οντότητα. Η

πρόκληση αυτή επιδεινώνεται από το γεγονός ότι τα βιογραφικά σημειώματα συχνά περιέχουν ειδική ορολογία, συντομογραφίες και ακρωνύμια, γεγονός που αυξάνει περαιτέρω την πιθανότητα ασάφειας (Bergersen et al., 2014).

Τα συνώνυμα προσθέτουν άλλο ένα επίπεδο πολυπλοκότητας στην NER για την ανάλυση βιογραφικού σημειώματος. Διαφορετικοί υποψήφιοι ενδέχεται να χρησιμοποιούν διαφορετικούς όρους για να περιγράψουν την ίδια δεξιότητα, ανάλογα με το υπόβαθρο, την εκπαίδευση ή την περιοχή τους. Για παράδειγμα, ένας υποψήφιος μπορεί να αναφέρει ως δεξιότητα την "ανάλυση δεδομένων", ενώ ένας άλλος μπορεί να χρησιμοποιεί την "ανάλυση δεδομένων". Και οι δύο όροι αναφέρονται στην ίδια γενική ικανότητα, αλλά ένα σύστημα NER πρέπει να είναι σε θέση να τους αναγνωρίζει και να τους αντιμετωπίζει ως ισοδύναμους. Ομοίως, όροι όπως "ανάπτυξη λογισμικού" και "μηχανική λογισμικού" μπορεί να χρησιμοποιούνται εναλλακτικά, αλλά ένα σύστημα NER πρέπει να κατανοήσει ότι αναφέρονται στην ίδια υποκείμενη ικανότητα (Gugnani & Misra, 2020).

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί τα συστήματα NER να ενσωματώνουν εξελιγμένες τεχνικές για το χειρισμό συνωνύμων και την επίλυση ασάφειας. Μια κοινή προσέγγιση είναι η χρήση προκαθορισμένων καταλόγων συνωνύμων και συναφών όρων, που συχνά αναφέρονται ως λεξικά ή λεξικά. Οι κατάλογοι αυτοί μπορεί να επιμελούνται χειροκίνητα ή να δημιουργούνται με τη χρήση αυτοματοποιημένων μεθόδων, όπως η ομαδοποίηση παρόμοιων όρων με βάση τη σημασιολογική τους ομοιότητα. Για παράδειγμα, ένα λεξικό για δεξιότητες προγραμματισμού μπορεί να περιλαμβάνει καταχωρίσεις όπως "προγραμματισμός Java", "ανάπτυξη Java", "κωδικοποίηση Java" και "σεναριογραφία Java", οι οποίες συνδέονται με μία μόνο κανονική δεξιότητα, τη "Java" (Bergersen et al., 2014).

Μια άλλη αποτελεσματική προσέγγιση περιλαμβάνει την αξιοποίηση μοντέλων μηχανικής μάθησης, ιδίως εκείνων που βασίζονται στη βαθιά μάθηση και σε μετασχηματιστές όπως οι BERT και GPT. Αυτά τα μοντέλα υπερέχουν στην κατανόηση του πλαισίου και στην καταγραφή των σημασιολογικών σχέσεων μεταξύ των λέξεων, καθιστώντας τα κατάλληλα για την επίλυση ασάφειας και την αναγνώριση συνωνύμων. Με την εκπαίδευση αυτών των μοντέλων σε μεγάλα, σχολιασμένα σύνολα δεδομένων που περιλαμβάνουν ποικίλα παραδείγματα του τρόπου με τον οποίο περιγράφονται οι δεξιότητες, τα συστήματα NER μπορούν να μάθουν να γενικεύουν και να αναγνωρίζουν με ακρίβεια ισοδύναμους όρους.

Για παράδειγμα, η BERT, με την αμφίδρομη αρχιτεκτονική της, μπορεί να καταλάβει ότι οι όροι "ανάπτυξη Java" και "προγραμματισμός Java" χρησιμοποιούνται εναλλακτικά σε διαφορετικά πλαίσια. Όταν τελειοποιηθεί σε ένα σώμα βιογραφικών σημειωμάτων και περιγραφών θέσεων εργασίας, η BERT μπορεί να μάθει να αναγνωρίζει αυτές τις παραλλαγές και να τις ταξινομεί ως την

ίδια δεξιότητα. Αυτή η ικανότητα βελτιώνει σημαντικά την ακρίβεια και την ευρωστία των συστημάτων NER, επιτρέποντάς τους να χειρίζονται τη μεταβλητότητα στον τρόπο περιγραφής των δεξιοτήτων σε διαφορετικά βιογραφικά σημειώματα (Zhao et al., 2015).

Συνάφεια

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι ένα κρίσιμο συστατικό της επεξεργασίας φυσικής γλώσσας (NLP) που χρησιμοποιείται στην ανάλυση βιογραφικών σημειωμάτων για την εξαγωγή σχετικών πληροφοριών, όπως δεξιότητες, τίτλοι θέσεων εργασίας και άλλα προσόντα. Μία από τις σημαντικές προκλήσεις στην ΕΑΑ για την ανάλυση βιογραφικών σημειωμάτων είναι η διασφάλιση της συνάφειας με το πλαίσιο, ιδίως στη διαφοροποίηση μεταξύ δεξιοτήτων και άλλων οντοτήτων (Sougandh et al., 2023). Αυτή η πρόκληση αναδεικνύεται από την ανάγκη διάκρισης μεταξύ όρων όπως "Python", οι οποίοι μπορεί να αναφέρονται τόσο σε μια γλώσσα προγραμματισμού όσο και σε έναν τύπο φιδιού. Η ακριβής ερμηνεία τέτοιων όρων στο πλαίσιο τους είναι απαραίτητη για την αποτελεσματική εξαγωγή δεξιοτήτων και την ακριβή επεξεργασία δεδομένων στην ανάλυση της αγοράς εργασίας (Zhang et al., 2019).

Η συνάφεια με το πλαίσιο στην ΕΑΑ περιλαμβάνει την κατανόηση του πλαισίου στο οποίο χρησιμοποιείται ένας όρος για την ακριβή ταξινόμησή του. Στην περίπτωση των βιογραφικών σημειωμάτων, αυτό σημαίνει ότι πρέπει να αναγνωρίσουμε ότι η λέξη "Python" στο πλαίσιο του βιογραφικού σημειώματος ενός προγραμματιστή λογισμικού πιθανώς αναφέρεται στη γλώσσα προγραμματισμού, ενώ σε ένα άλλο πλαίσιο μπορεί να αναφέρεται στο ερπετό. Αυτή η εργασία απαιτεί εξελιγμένα μοντέλα που μπορούν να αναλύσουν και να ερμηνεύσουν το περιβάλλον κείμενο για να καθορίσουν τη σωστή ταξινόμηση κάθε όρου (Chifu et al., 2017).

Τα παραδοσιακά συστήματα που βασίζονται σε κανόνες δυσκολεύονται να αντιμετωπίσουν αυτή την πρόκληση, επειδή βασίζονται σε προκαθορισμένα μοτίβα και λέξεις-κλειδιά, τα οποία δεν έχουν την ευελιξία να ερμηνεύσουν με ακρίβεια το πλαίσιο. Για παράδειγμα, ένα σύστημα βασισμένο σε κανόνες μπορεί να επισημάνει κάθε περίπτωση της λέξης "Python" ως γλώσσα προγραμματισμού, αποτυγχάνοντας να λάβει υπόψη τις περιπτώσεις όπου αναφέρεται στο φίδι. Αυτή η έλλειψη επίγνωσης του πλαισίου μπορεί να οδηγήσει σε σημαντικές ανακρίβειες στην εξαγωγή δεδομένων, υπονομεύοντας την αποτελεσματικότητα του συστήματος NER (Sougandh et al., 2023).

Οι προσεγγίσεις μηχανικής μάθησης, ιδίως εκείνες που περιλαμβάνουν βαθιά μάθηση και μοντέλα μετασχηματιστών όπως το BERT (Bidirectional Encoder Representations from Transformers) και το GPT (Generative Pre-trained Transformer), προσφέρουν πιο προηγμένες λύσεις. Αυτά τα μοντέλα έχουν

σχεδιαστεί για να κατανοούν και να επεξεργάζονται το πλαίσιο, γεγονός που τα καθιστά κατάλληλα για το χειρισμό των διαφοροποιημένων διακρίσεων που απαιτούνται στην NER για τη σύνταξη συνόψεων (Chifu et al., 2017).

Το BERT, για παράδειγμα, επεξεργάζεται το κείμενο αμφίδρομα, δηλαδή εξετάζει το περιεχόμενο τόσο από τα αριστερά όσο και από τα δεξιά κάθε λέξης. Αυτή η ικανότητα επιτρέπει στο BERT να κατανοεί τις αποχρώσεις της γλώσσας και να κάνει πιο ακριβείς προβλέψεις σχετικά με τη σημασία μιας λέξης με βάση το περιβάλλον της. Στο παράδειγμα της λέξης "Python", η BERT μπορεί να αναλύσει τις λέξεις γύρω της για να καθορίσει αν ο όρος χρησιμοποιείται σε τεχνικό πλαίσιο που σχετίζεται με την ανάπτυξη λογισμικού ή σε βιολογικό πλαίσιο που αναφέρεται στο φίδι (Gugnani & Misra, 2020).

Για την αποτελεσματική εκπαίδευση αυτών των μοντέλων, απαιτούνται μεγάλα σύνολα δεδομένων με σχόλια, όπου οι όροι επισημαίνονται σύμφωνα με τις σωστές οντότητές τους με βάση τα συμφραζόμενα. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει να αναγνωρίζει μοτίβα και σχέσεις στα δεδομένα που υποδεικνύουν τη σωστή ταξινόμηση κάθε όρου. Αφού εκπαιδευτούν, τα μοντέλα αυτά μπορούν να εφαρμόσουν την κατανόηση που έχουν μάθει σε νέο, αθέατο κείμενο, αναγνωρίζοντας και ταξινομώντας με ακρίβεια οντότητες με βάση τα συμφραζόμενα (Sougandh et al., 2023).

Τα μοντέλα GPT, και ιδιαίτερα το GPT-3, παρέχουν επίσης αξιόπιστες λύσεις για τη συνάφεια με το περιβάλλον στη NEP. Η τεράστια κλίμακα και οι δυνατότητες βαθιάς μάθησης του GPT-3 του επιτρέπουν να παράγει αναπαραστάσεις κειμένου με υψηλή επίγνωση του πλαισίου. Αξιοποιώντας την εκτεταμένη εκπαίδευσή του σε ποικίλα σύνολα δεδομένων, το GPT-3 μπορεί να διακρίνει το επιδιωκόμενο νόημα διφορούμενων όρων όπως "Python" λαμβάνοντας υπόψη το ευρύτερο πλαίσιο στο οποίο εμφανίζονται (Chifu et al., 2017).

Η εφαρμογή αυτών των προηγμένων μοντέλων στην ανάλυση βιογραφικών σημειωμάτων έχει σημαντικές επιπτώσεις στην ανάλυση της αγοράς εργασίας. Με την ακριβή διάκριση μεταξύ δεξιοτήτων και άλλων οντοτήτων, τα συστήματα NER μπορούν να διασφαλίσουν ότι τα βιογραφικά σημειώματα αναλύονται με μεγάλη ακρίβεια, εξάγοντας μόνο τις σχετικές πληροφορίες που απαιτούνται για την αντιστοίχιση θέσεων εργασίας και την αξιολόγηση των υποψηφίων. Για παράδειγμα, η αναγνώριση της "Python" ως γλώσσας προγραμματισμού σε ένα βιογραφικό σημείωμα επιτρέπει στους υπεύθυνους προσλήψεων να αξιολογούν με ακρίβεια τις τεχνικές δεξιότητες ενός υποψηφίου και να τον αντιστοιχίζουν με τις κατάλληλες ευκαιρίες απασχόλησης (Zhang et al., 2019).

Ποιότητα δεδομένων και θόρυβος

Η αναγνώριση ονομαστικών οντοτήτων (NER) στο πλαίσιο της ανάλυσης βιογραφικών σημειωμάτων είναι ζωτικής σημασίας για την ακριβή εξαγωγή

σχετικών πληροφοριών από βιογραφικά σημειώματα, όπως δεξιότητες, τίτλοι εργασίας και προσόντα. Ωστόσο, μία από τις σημαντικότερες προκλήσεις σε αυτή τη διαδικασία είναι η αντιμετώπιση της ποιότητας των δεδομένων και του θορύβου. Τα βιογραφικά σημειώματα συχνά περιέχουν ελλιπή ή κακώς μορφοποιημένα δεδομένα, γεγονός που μπορεί να παρεμποδίσει σημαντικά την απόδοση των συστημάτων NER. Ζητήματα όπως ορθογραφικά λάθη, ασυνεπής μορφοποίηση και μη συμβατικές δομές βιογραφικών σημειωμάτων είναι συνηθισμένα, παρουσιάζοντας σημαντικά εμπόδια στην ακριβή εξαγωγή και ταξινόμηση οντοτήτων (Zhang et al., 2019).

Τα ελλιπή δεδομένα αποτελούν συχνό πρόβλημα στην ανάλυση βιογραφικών σημειωμάτων. Οι υποψήφιοι μπορεί να παραλείπουν ορισμένες λεπτομέρειες, όπως συγκεκριμένους τίτλους εργασίας, ημερομηνίες απασχόλησης ή εκπαιδευτικά προσόντα, είτε σκόπιμα είτε ακούσια. Αυτή η έλλειψη πληροφοριών αποτελεί πρόκληση για τα συστήματα NER, τα οποία βασίζονται σε ενδείξεις του περιβάλλοντος για τον ακριβή εντοπισμό και την ταξινόμηση οντοτήτων. Για παράδειγμα, ένα βιογραφικό σημείωμα μπορεί να αναφέρει μια θέση ως "Διευθυντής" χωρίς να προσδιορίζει την ακριβή φύση του διευθυντικού ρόλου ή το τμήμα που εμπλέκεται. Χωρίς πρόσθετο περιεχόμενο, το σύστημα NER μπορεί να δυσκολευτεί να ταξινομήσει με ακρίβεια αυτόν τον τίτλο εργασίας, οδηγώντας σε πιθανές αναντιστοιχίες στην ευθυγράμμιση εργασίας-υποψηφίου (Mulalic, 2019).

Τα κακώς μορφοποιημένα δεδομένα είναι μια άλλη σημαντική πρόκληση. Τα βιογραφικά σημειώματα μπορούν να υποβληθούν σε διάφορες μορφές, όπως PDF, Word και απλό κείμενο, η καθεμία με τα δικά της προβλήματα μορφοποίησης. Ακόμη και εντός της ίδιας μορφής, οι υποψήφιοι μπορεί να χρησιμοποιούν διαφορετικά στυλ, διατάξεις και δομές, γεγονός που καθιστά δύσκολο για τα συστήματα NER να εξάγουν με συνέπεια τις σχετικές πληροφορίες. Για παράδειγμα, ορισμένοι υποψήφιοι μπορεί να χρησιμοποιούν πίνακες για να οργανώσουν την εργασιακή τους εμπειρία, ενώ άλλοι μπορεί να χρησιμοποιούν κουκκίδες ή συνεχές κείμενο. Αυτές οι παραλλαγές μπορούν να προκαλέσουν σύγχυση στα συστήματα NER, τα οποία μπορεί να χάσουν ή να ταξινομήσουν εσφαλμένα οντότητες λόγω της ασυνεπούς μορφοποίησης.

Τα ορθογραφικά και τυπογραφικά λάθη είναι επίσης διαδεδομένα στα βιογραφικά σημειώματα και μπορούν να επηρεάσουν σοβαρά την ακρίβεια των συστημάτων NER. Οι ανορθόγραφες λέξεις ενδέχεται να μην ταιριάζουν με τους προκαθορισμένους καταλόγους οντοτήτων ή τα δεδομένα εκπαίδευσης που χρησιμοποιούνται από το μοντέλο NER, οδηγώντας σε εσφαλμένη ή αποτυχημένη αναγνώριση οντοτήτων. Για παράδειγμα, εάν ένα βιογραφικό σημείωμα περιέχει

"Pythn" αντί για "Python", ένα σύστημα NER μπορεί να μην το αναγνωρίσει ως αναφορά στη γλώσσα προγραμματισμού, χάνοντας έτσι μια κρίσιμη δεξιότητα.

Ο χειρισμός αντισυμβατικών μορφών και δημιουργικών σχεδίων βιογραφικού σημειώματος προσθέτει άλλο ένα επίπεδο πολυπλοκότητας. Οι υποψήφιοι χρησιμοποιούν συχνά μοναδικές διατάξεις για να κάνουν το βιογραφικό τους να ξεχωρίζει, όπως η χρήση γραφικών, infographics ή προσαρμοσμένων γραμματοσειρών. Ενώ αυτά τα σχέδια μπορεί να είναι οπτικά ελκυστικά, δημιουργούν σημαντικές προκλήσεις για τα συστήματα NER, τα οποία συνήθως εκπαιδεύονται σε πιο τυποποιημένες διατάξεις κειμένου. Η εξαγωγή ουσιαστικών δεδομένων από αυτές τις μη συμβατικές μορφές απαιτεί προηγμένες τεχνικές προεπεξεργασίας και ισχυρά μοντέλα NER ικανά να προσαρμόζονται σε διάφορα στυλ παρουσίασης (Mulalic, 2019).

Για τη διαχείριση αυτών των προκλήσεων μπορούν να χρησιμοποιηθούν διάφορες στρατηγικές. Οι προηγμένες τεχνικές προεπεξεργασίας είναι απαραίτητες για τη βελτίωση της ποιότητας των δεδομένων πριν από την εφαρμογή της NER. Αυτό περιλαμβάνει την κανονικοποίηση του κειμένου με τη μετατροπή διαφορετικών μορφών σε μια συνεπή δομή, τη διόρθωση ορθογραφικών σφαλμάτων με αλγόριθμους ορθογραφικού ελέγχου και τον χειρισμό διαφόρων μορφών διάταξης μέσω εξελεγμένων μεθόδων ανάλυσης. Για παράδειγμα, η τεχνολογία OCR (οπτική αναγνώριση χαρακτήρων) μπορεί να χρησιμοποιηθεί για τη μετατροπή εγγράφων PDF σε κείμενο αναγνώσιμο από μηχανήματα, ενώ οι αλγόριθμοι ορθογραφικού ελέγχου και κανονικοποίησης κειμένου μπορούν να διορθώσουν συνήθη προβλήματα ορθογραφίας και μορφοποίησης (Todorovic et al., 2010).

1.4.3. Μελέτες περιπτώσεων και παραδείγματα

Kenexa Talent Acquisition Suite της IBM

Η αναγνώριση ονομαστικών οντοτήτων (NER) έχει γίνει μια μετασχηματιστική τεχνολογία στον τομέα των προσλήψεων, βελτιώνοντας σημαντικά την αποτελεσματικότητα και την ακρίβεια των διαδικασιών ελέγχου βιογραφικών σημειωμάτων. Ένα χαρακτηριστικό παράδειγμα της εφαρμογής της NER στην πρόσληψη μπορεί να δει κανείς στην εφαρμογή της IBM στο πλαίσιο της σουίτας Kenexa Talent Acquisition Suite. Αυτή η μελέτη περίπτωσης καταδεικνύει τον τρόπο με τον οποίο η NER μπορεί να βελτιώσει τις ροές εργασίας πρόσληψης, να βελτιώσει την αντιστοιχία υποψηφίων και τελικά να οδηγήσει σε καλύτερα αποτελέσματα προσλήψεων (Attupuram et al., 2015).

Το Kenexa Talent Acquisition Suite της IBM είναι μια ολοκληρωμένη πλατφόρμα προσλήψεων που έχει σχεδιαστεί για την υποστήριξη ολοκληρωμένων

διαδικασιών πρόσληψης. Για την περαιτέρω ενίσχυση των δυνατοτήτων της πλατφόρμας, η IBM ενσωμάτωσε ένα εξελιγμένο εργαλείο ελέγχου βιογραφικών σημειωμάτων με βάση το NER. Το εργαλείο αυτό αξιοποιεί αλγορίθμους NER για την αυτόματη εξαγωγή κρίσιμων πληροφοριών από βιογραφικά σημειώματα, όπως δεξιότητες, εκπαιδευτικά προσόντα και εργασιακή εμπειρία. Με τη μετατροπή των αδόμητων δεδομένων βιογραφικών σημειωμάτων σε δομημένες πληροφορίες, το εργαλείο NER επιτρέπει την αποτελεσματικότερη και ακριβέστερη επεξεργασία των υποψηφίων (Saad et al., 2021).

Η εφαρμογή του NER στο πλαίσιο της σουίτας προσέλκυσης ταλέντων της Kenexa έχει αποφέρει σημαντικά οφέλη. Μία από τις σημαντικότερες επιπτώσεις ήταν η μείωση του χρόνου πρόσληψης, ο οποίος μειώθηκε κατά 30%. Η μείωση αυτή οφείλεται κυρίως στην αυτοματοποίηση της αρχικής διαδικασίας διαλογής βιογραφικών σημειωμάτων, η οποία παραδοσιακά περιλαμβάνει σημαντική χειρωνακτική προσπάθεια. Το εργαλείο NER επεξεργάζεται γρήγορα μεγάλους όγκους βιογραφικών σημειωμάτων, εντοπίζοντας και κατηγοριοποιώντας σχετικές πληροφορίες που διαφορετικά θα απαιτούσαν εκτενή ανθρώπινη εξέταση. Ως αποτέλεσμα, οι υπεύθυνοι προσλήψεων μπορούν να επικεντρώσουν τις προσπάθειές τους στην επαφή με τους πιο υποσχόμενους υποψηφίους, επιταχύνοντας τη συνολική διαδικασία πρόσληψης (Attururam et al., 2015).

Εκτός από τη βελτίωση της αποτελεσματικότητας, το εργαλείο NER έχει επίσης βελτιώσει την ακρίβεια της αντιστοίχισης των υποψηφίων. Με την ακριβή εξαγωγή και κατηγοριοποίηση των βασικών πληροφοριών βιογραφικού σημειώματος, το εργαλείο διασφαλίζει ότι οι υποψήφιοι αξιολογούνται με βάση τα σχετικά κριτήρια. Αυτή η ακριβής αντιστοίχιση οδηγεί σε υψηλότερα ποσοστά ικανοποίησης μεταξύ των υπευθύνων προσλήψεων, οι οποίοι μπορούν να εντοπίζουν ευκολότερα τους υποψηφίους που πληρούν τις συγκεκριμένες απαιτήσεις των θέσεων εργασίας. Η βελτιωμένη ακρίβεια συμβάλλει επίσης στην καλύτερη εμπειρία των υποψηφίων, καθώς οι υποψήφιοι είναι πιο πιθανό να εξεταστούν για ρόλους που ταιριάζουν στενά με τις δεξιότητες και τις εμπειρίες τους (Tran et al., 2022).

Η ενσωμάτωση της NER στη διαδικασία προσλήψεων της IBM δεν ήταν χωρίς προκλήσεις. Μία από τις πρωταρχικές προκλήσεις ήταν η διασφάλιση της ακρίβειας του εργαλείου στις διάφορες μορφές και δομές των βιογραφικών σημειωμάτων που υπέβαλαν οι υποψήφιοι. Τα βιογραφικά σημειώματα μπορεί να ποικίλλουν σε μεγάλο βαθμό ως προς την παρουσίασή τους, από απλά έγγραφα κειμένου έως σύνθετα σχέδια με γραφικά και στήλες. Για την αντιμετώπιση αυτού του προβλήματος, η IBM χρησιμοποίησε προηγμένες τεχνικές προεπεξεργασίας για την κανονικοποίηση των μορφών βιογραφικών σημειωμάτων, καθιστώντας τα συμβατά με το εργαλείο NER. Επιπλέον, οι

αλγόριθμοι NER τελειοποιήθηκαν με τη χρήση εκτεταμένων σχολιασμένων συνόλων δεδομένων που αντανάκλυσαν τη μεταβλητότητα του περιεχομένου των βιογραφικών σημειωμάτων (Attupuram et al., 2015).

Το εργαλείο NER της IBM χρησιμοποιεί έναν συνδυασμό τυχαίων πεδίων υπό συνθήκη (CRF) και δικτύων αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (BiLSTM) για την αναγνώριση οντοτήτων. Τα CRF χρησιμοποιούνται για την ικανότητά τους να μοντελοποιούν τις εξαρτήσεις μεταξύ των ετικετών σε μια ακολουθία, η οποία είναι ζωτικής σημασίας για την ακριβή αναγνώριση οντοτήτων στο πλαίσιο ενός βιογραφικού σημειώματος. Για παράδειγμα, τα CRF μπορούν να διακρίνουν αποτελεσματικά μεταξύ ενός τίτλου εργασίας και μιας δεξιοτήτας με βάση το περιβάλλον κείμενο. Τα δίκτυα BiLSTM συμπληρώνουν αυτό παρέχοντας βαθιά κατανόηση του πλαισίου μέσω της ικανότητάς τους να επεξεργάζονται κείμενο τόσο προς τα εμπρός όσο και προς τα πίσω. Αυτή η αμφίδρομη επεξεργασία επιτρέπει στο εργαλείο NER να συλλάβει τις αποχρώσεις των σχέσεων μεταξύ των λέξεων, βελτιώνοντας τη συνολική ακρίβεια της αναγνώρισης οντοτήτων (Saad et al., 2021).

Έρευνα του Πανεπιστημίου Stanford για το BiLSTM-CRF για NER

Τα τελευταία χρόνια, η ακαδημαϊκή έρευνα έχει επικεντρωθεί όλο και περισσότερο στην εφαρμογή προηγμένων μοντέλων μηχανικής μάθησης για την αναγνώριση ονομαστικών οντοτήτων (NER) στην ανάλυση της αγοράς εργασίας. Μια αξιοσημείωτη μελέτη από το Πανεπιστήμιο του Στάνφορντ αποτελεί παράδειγμα αυτής της τάσης διερευνώντας την αποτελεσματικότητα των δικτύων αμφίδρομης μακράς βραχυπρόθεσμης μνήμης σε συνδυασμό με τυχαία πεδία υπό συνθήκη (BiLSTM-CRF) για την εξαγωγή δεξιοτήτων από βιογραφικά σημειώματα. Η εν λόγω έρευνα ανέδειξε την ανώτερη απόδοση των μοντέλων BiLSTM-CRF σε σύγκριση με τις παραδοσιακές μεθόδους NER, ιδίως όσον αφορά την ακρίβεια και την ανάκληση, προσφέροντας σημαντικές επιπτώσεις για τον κλάδο των ανθρώπινων πόρων (Tran et al., 2022).

Η μελέτη που διεξήχθη από το Πανεπιστήμιο του Στάνφορντ αποσκοπούσε στην αντιμετώπιση των προκλήσεων που σχετίζονται με την εξαγωγή σύνθετων οντοτήτων δεξιοτήτων από αδόμητο κείμενο βιογραφικών σημειωμάτων. Οι παραδοσιακές προσεγγίσεις NER συχνά αντιμετώπιζαν τη μεταβλητότητα και την εξάρτηση από το πλαίσιο του περιεχομένου του βιογραφικού σημειώματος, οδηγώντας σε ανακρίβειες στην εξαγωγή δεξιοτήτων. Για να ξεπεράσουν αυτούς τους περιορισμούς, οι ερευνητές χρησιμοποίησαν μοντέλα BiLSTM-CRF, αξιοποιώντας τα πλεονεκτήματα τόσο των BiLSTM όσο και των CRF για να ενισχύσουν την ακρίβεια της αναγνώρισης οντοτήτων (Jagwani et al., 2023).

Τα BiLSTM είναι ένας τύπος επαναλαμβανόμενου νευρωνικού δικτύου (RNN) που επεξεργάζεται ακολουθίες κειμένου τόσο προς τα εμπρός όσο και προς τα πίσω,

συλλαμβάνοντας τα συμφραζόμενα τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις. Αυτή η αμφίδρομη ικανότητα επιτρέπει στα BiLSTM να κατανοούν τις αποχρώσεις των σχέσεων μεταξύ των λέξεων, πράγμα που είναι ζωτικής σημασίας για τον ακριβή εντοπισμό οντοτήτων σε σύνθετα κείμενα. Οι CRF, από την άλλη πλευρά, είναι στατιστικά μοντέλα που υπερέχουν στην επισήμανση ακολουθιών λαμβάνοντας υπόψη τις εξαρτήσεις μεταξύ των ετικετών. Συνδυάζοντας τα BiLSTMs με τα CRFs, το μοντέλο μπορεί να αξιοποιήσει την κατανόηση των συμφραζομένων των BiLSTMs και τις δυνατότητες δομημένης πρόβλεψης των CRFs, με αποτέλεσμα την ακριβέστερη και πιο ισχυρή αναγνώριση οντοτήτων (Ushio & Camacho-Collados, 2022).

Η μεθοδολογία της μελέτης περιελάμβανε την εκπαίδευση του μοντέλου BiLSTM-CRF σε ένα μεγάλο, σχολιασμένο σύνολο δεδομένων βιογραφικών σημειωμάτων. Το σύνολο δεδομένων περιλάμβανε ένα ευρύ φάσμα βιογραφικών με χειροκίνητα επισημειωμένες οντότητες, όπως δεξιότητες, τίτλους εργασίας και εκπαιδευτικά προσόντα. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο έμαθε να αναγνωρίζει μοτίβα και σχέσεις μέσα στο κείμενο που υποδεικνυαν την παρουσία αυτών των οντοτήτων. Οι ερευνητές χρησιμοποίησαν τεχνικές διασταυρούμενης επικύρωσης για να διασφαλίσουν την ευρωστία του μοντέλου και να αποφύγουν την υπερβολική προσαρμογή (Jagwani et al., 2023).

spraCy για NER σε εφαρμογές HR

Τα έργα και τα εργαλεία ανοικτού κώδικα έχουν διαδραματίσει μετασχηματιστικό ρόλο στην ανάπτυξη εφαρμογών αναγνώρισης ονομαστικών οντοτήτων (NER), ιδίως στον τομέα της τεχνολογίας του ανθρώπινου δυναμικού (HR). Μεταξύ αυτών, το *spraCy* ξεχωρίζει ως μια ευρέως υιοθετημένη βιβλιοθήκη NLP που έχει γίνει ακρογωνιαίος λίθος για την κατασκευή εφαρμογών NER. Οι ισχυρές δυνατότητες του *spraCy*, το φιλικό προς το χρήστη περιβάλλον και η ευελιξία του το έχουν καταστήσει μια ελκυστική επιλογή για τις νεοσύστατες εταιρείες τεχνολογίας HR που στοχεύουν στην ανάπτυξη προσαρμοσμένων εργαλείων NER για την ανάλυση βιογραφικών σημειωμάτων και την αντιστοίχιση θέσεων εργασίας. Αυτή η μελέτη περίπτωσης διερευνά τον τρόπο με τον οποίο το *spraCy* έχει χρησιμοποιηθεί σε εφαρμογές HR, αναδεικνύοντας τον αντίκτυπό του στη βελτίωση της ακρίβειας και της αποτελεσματικότητας της ανάπτυξης (Ushio & Camacho-Collados, 2022).

Το *spraCy* είναι μια βιβλιοθήκη NLP ανοικτού κώδικα που έχει σχεδιαστεί για να παρέχει προηγμένες λειτουργίες για την επεξεργασία και ανάλυση κειμένου. Η δημοτικότητά της πηγάζει από το ολοκληρωμένο σύνολο χαρακτηριστικών της, συμπεριλαμβανομένων προ-εκπαιδευμένων μοντέλων, ισχυρών δυνατοτήτων ενσωμάτωσης και εκτεταμένων επιλογών προσαρμογής. Αυτά τα χαρακτηριστικά καθιστούν το *spraCy* ιδιαίτερα κατάλληλο για εργασίες NER σε

εφαρμογές ανθρώπινου δυναμικού, όπου η ανάγκη ακριβούς εξαγωγής και ταξινόμησης πληροφοριών από βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας είναι υψίστης σημασίας (Chifu et al., 2017).

Ένα από τα βασικά χαρακτηριστικά του spaCy που το καθιστά ιδανικό για εφαρμογές HR είναι τα προ-εκπαιδευμένα μοντέλα του. Αυτά τα μοντέλα έχουν εκπαιδευτεί σε μεγάλα σώματα δεδομένων και είναι ικανά να αναγνωρίζουν ένα ευρύ φάσμα οντοτήτων από το κουτί. Για τις νεοσύστατες εταιρείες τεχνολογίας HR, αυτό σημαίνει ότι μπορούν να αξιοποιήσουν τα προ-εκπαιδευμένα μοντέλα του spaCy για να αναπτύξουν και να αναπτύξουν γρήγορα εφαρμογές NER χωρίς την ανάγκη εκτεταμένης εκπαίδευσης σε προσαρμοσμένα σύνολα δεδομένων. Αυτό μειώνει σημαντικά τον χρόνο ανάπτυξης και την προσπάθεια που απαιτείται για τη δημιουργία αποτελεσματικών εργαλείων ανάλυσης βιογραφικών σημειωμάτων (Vinic, 2023). Για παράδειγμα, μια νεοσύστατη επιχείρηση που αναπτύσσει μια πλατφόρμα αντιστοίχισης θέσεων εργασίας μπορεί να χρησιμοποιήσει το προ-εκπαιδευμένο μοντέλο NER της spaCy για να εξάγει οντότητες όπως τίτλους θέσεων εργασίας, δεξιότητες και ονόματα εταιρειών από βιογραφικά σημειώματα, επιτρέποντας την ακριβέστερη αντιστοίχιση υποψηφίων και θέσεων εργασίας (Walford-Wright & Scott-Jackson, 2018).

Οι δυνατότητες ενσωμάτωσης του spaCy ενισχύουν περαιτέρω την καταλληλότητά του για εφαρμογές ανθρώπινου δυναμικού. Το spaCy μπορεί εύκολα να ενσωματωθεί με άλλα εργαλεία και πλαίσια, επιτρέποντας στις νεοσύστατες επιχειρήσεις να δημιουργήσουν ολοκληρωμένες λύσεις ανθρώπινου δυναμικού. Για παράδειγμα, το spaCy μπορεί να συνδυαστεί με βιβλιοθήκες μηχανικής μάθησης όπως το TensorFlow ή το PyTorch για τη δημιουργία πιο εξελιγμένων μοντέλων που βελτιώνουν την ακρίβεια αναγνώρισης οντοτήτων. Επιπλέον, η συμβατότητα του spaCy με άλλα εργαλεία επεξεργασίας δεδομένων, όπως το pandas για χειρισμό δεδομένων και το scikit-learn για μηχανική μάθηση, επιτρέπει απρόσκοπτες ροές εργασίας δεδομένων. Αυτή η δυνατότητα ενσωμάτωσης επιτρέπει στις νεοσύστατες εταιρείες τεχνολογίας HR να δημιουργούν ολοκληρωμένες λύσεις που βελτιώνουν τη διαδικασία πρόσληψης, από την εξαγωγή δεδομένων έως την αξιολόγηση των υποψηφίων (Chifu et al., 2017).

Η προσαρμογή είναι ένα άλλο σημαντικό πλεονέκτημα του spaCy. Ενώ τα προ-εκπαιδευμένα μοντέλα παρέχουν ένα ισχυρό σημείο εκκίνησης, το spaCy επιτρέπει εκτεταμένη προσαρμογή για την κάλυψη συγκεκριμένων αναγκών. Οι νεοσύστατες επιχειρήσεις μπορούν να προσαρμόσουν τα προ-εκπαιδευμένα μοντέλα στα δικά τους σύνολα δεδομένων για να βελτιώσουν τις επιδόσεις σε εργασίες συγκεκριμένου τομέα. Για παράδειγμα, μια νεοσύστατη επιχείρηση που

επικεντρώνεται στον κλάδο της τεχνολογίας θα μπορούσε να προσαρμόσει το μοντέλο NER του spaCy χρησιμοποιώντας ένα σύνολο δεδομένων με βιογραφικά σημειώματα και περιγραφές θέσεων εργασίας στον κλάδο της τεχνολογίας, βελτιώνοντας την ικανότητα του μοντέλου να αναγνωρίζει όρους και δεξιότητες που αφορούν συγκεκριμένο κλάδο. Αυτή η προσαρμογή διασφαλίζει ότι τα εργαλεία NER είναι εξαιρετικά συναφή και ακριβή για το συγκεκριμένο πλαίσιο στο οποίο χρησιμοποιούνται (Ushio & Camacho-Collados, 2022).

Η χρήση της NER από το LinkedIn για τη βελτίωση της μηχανής σύστασης θέσεων εργασίας

Η αναγνώριση ονομαστικών οντοτήτων (NER) έχει βρει εκτεταμένη εφαρμογή στη βελτίωση της αποτελεσματικότητας και της ακρίβειας των μηχανών σύστασης θέσεων εργασίας, με χαρακτηριστικό παράδειγμα τη χρήση προηγμένων τεχνικών NER από το LinkedIn. Το LinkedIn αξιοποιεί αυτές τις τεχνικές για να εξάγει κρίσιμες πληροφορίες από τα προφίλ των χρηστών και τις αγγελίες εργασίας, όπως δεξιότητες, τίτλους θέσεων εργασίας και άλλες σχετικές οντότητες. Αυτά τα εξαγόμενα δεδομένα είναι αναπόσπαστο στοιχείο για τη βελτίωση της μηχανής σύστασης θέσεων εργασίας, παρέχοντας έτσι στους χρήστες ακριβέστερες και πιο σχετικές προτάσεις θέσεων εργασίας. Η εφαρμογή της NER στο LinkedIn έχει ενισχύσει σημαντικά τη δέσμευση και την ικανοποίηση των χρηστών, αναδεικνύοντας τις μετασχηματιστικές δυνατότητες αυτής της τεχνολογίας στην ανάλυση της αγοράς εργασίας (Walford-Wright & Scott-Jackson, 2018).

Η μηχανή σύστασης θέσεων εργασίας του LinkedIn βασίζεται σε μεγάλο βαθμό στην ακριβή εξαγωγή και ταξινόμηση οντοτήτων από τεράστιους όγκους δεδομένων κειμένου. Τα προφίλ στο LinkedIn περιέχουν πληθώρα πληροφοριών, όπως το επαγγελματικό ιστορικό των χρηστών, τις δεξιότητες, το εκπαιδευτικό υπόβαθρο και τις επαγγελματικές εγκρίσεις. Ομοίως, οι αγγελίες θέσεων εργασίας περιγράφουν λεπτομερώς τα απαιτούμενα προσόντα, τις αρμοδιότητες και τις πληροφορίες της εταιρείας. Χρησιμοποιώντας εξελιγμένες τεχνικές NER, το LinkedIn μπορεί να αναλύει και να κατηγοριοποιεί συστηματικά αυτές τις πληροφορίες, δημιουργώντας δομημένα σύνολα δεδομένων που οδηγούν τους αλγορίθμους συστάσεων (Palshikar et al., 2018).

Μία από τις προηγμένες τεχνικές NER που χρησιμοποιεί το LinkedIn περιλαμβάνει μοντέλα μετασχηματιστών, ιδίως το BERT (Bidirectional Encoder Representations from Transformers). Η αμφίδρομη αρχιτεκτονική του BERT του επιτρέπει να λαμβάνει υπόψη του το πλαίσιο τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις, πράγμα που είναι ζωτικής σημασίας για την ακριβή ερμηνεία της διαφοροποιημένης και ποικίλης γλώσσας που συναντάται στα βιογραφικά σημειώματα και τις περιγραφές θέσεων εργασίας. Για παράδειγμα, η

BERT μπορεί να διακρίνει μεταξύ της "Java" ως γλώσσας προγραμματισμού και της "Java" ως γεωγραφικής τοποθεσίας με βάση το περιβάλλον. Αυτό το επίπεδο κατανόησης των συμφραζομένων είναι απαραίτητο για την εξαγωγή ακριβών και σχετικών οντοτήτων από περιεχόμενο που δημιουργείται από χρήστες και καταχωρίσεις θέσεων εργασίας (Walford-Wright & Scott-Jackson, 2018).

Η χρήση μοντέλων μετασχηματιστών όπως το BERT επιτρέπει στο LinkedIn να επιτυγχάνει υψηλή ακρίβεια στην αναγνώριση οντοτήτων, γεγονός που επηρεάζει άμεσα την ποιότητα των συστάσεων εργασίας. Αναγνωρίζοντας με ακρίβεια τις δεξιότητες και τους τίτλους εργασίας από τα προφίλ και τις δημοσιεύσεις, η μηχανή συστάσεων του LinkedIn μπορεί να αντιστοιχίσει τους χρήστες με ευκαιρίες εργασίας που ταιριάζουν στενά με τα προσόντα και τα επαγγελματικά τους ενδιαφέροντα. Αυτή η ακριβής αντιστοίχιση όχι μόνο βελτιώνει την εμπειρία του χρήστη αλλά και αυξάνει την πιθανότητα επιτυχούς τοποθέτησης σε θέσεις εργασίας, προς όφελος τόσο των αναζητούντων εργασίας όσο και των εργοδοτών (Hu, 2023).

Η εφαρμογή της NER στην κλίμακα που απαιτεί το LinkedIn παρουσιάζει αρκετές προκλήσεις, ιδίως όσον αφορά τον χειρισμό δεδομένων μεγάλης κλίμακας και τη διασφάλιση της ακρίβειας και της ιδιωτικότητας των δεδομένων. Το LinkedIn επεξεργάζεται καθημερινά έναν τεράστιο όγκο δεδομένων, με εκατομμύρια προφίλ χρηστών και δημοσιεύσεις θέσεων εργασίας να ενημερώνονται και να δημιουργούνται. Η αποτελεσματική διαχείριση αυτών των δεδομένων απαιτεί ισχυρή υποδομή και βελτιστοποιημένους αλγορίθμους με δυνατότητα επεξεργασίας σε πραγματικό χρόνο. Το LinkedIn αντιμετωπίζει αυτές τις προκλήσεις αξιοποιώντας κατανομημένα υπολογιστικά πλαίσια και λύσεις που βασίζονται στο cloud και διευκολύνουν την κλιμακούμενη επεξεργασία και αποθήκευση δεδομένων (Palshikar et al., 2018).

2. Η κατάσταση της τέχνης

2.1. Ανάλυση παραδοσιακών δεδομένων της αγοράς εργασίας

2.1.1 Ιστορικές προσεγγίσεις για την ανάλυση της αγοράς εργασίας

Έρευνες και ερωτηματολόγια

Η Έρευνα Τρέχοντος Πληθυσμού (CPS) του Γραφείου Στατιστικής Εργασίας των ΗΠΑ (BLS) αποτελεί ακρογωνιαίο λίθο στην παραδοσιακή ανάλυση των δεδομένων της αγοράς εργασίας, παρέχοντας ολοκληρωμένα και έγκαιρα στοιχεία για την απασχόληση, την ανεργία και άλλους κρίσιμους δείκτες της αγοράς εργασίας. Η CPS διεξάγεται μηνιαίως και αποτελεί μία από τις κύριες πηγές πληροφοριών για την αγορά εργασίας στις Ηνωμένες Πολιτείες, επηρεάζοντας σημαντικά την κυβερνητική πολιτική, την οικονομική έρευνα και την κατανόηση των τάσεων της αγοράς εργασίας από το κοινό (Groves & Peytcheva, 2008).

Η CPS χρησιμοποιεί μια αυστηρή μεθοδολογία για να διασφαλίσει την ακρίβεια και την αξιοπιστία των δεδομένων της. Η έρευνα χρησιμοποιεί μια τεχνική στρωματοποιημένης δειγματοληψίας, επιλέγοντας νοικοκυριά και από τις 50 πολιτείες και την Περιφέρεια της Κολούμπια. Η προσέγγιση αυτή εξασφαλίζει ένα αντιπροσωπευτικό δείγμα που αποτυπώνει την ποικιλομορφία του πληθυσμού των ΗΠΑ. Κάθε μήνα, ερευνώνται περίπου 60.000 επιλέξιμα νοικοκυριά, παρέχοντας ένα μεγάλο μέγεθος δείγματος που ενισχύει τη στατιστική αξιοπιστία των ευρημάτων. Τα νοικοκυριά ερωτώνται για τέσσερις συνεχόμενους μήνες, μένουν εκτός δείγματος για οκτώ μήνες και στη συνέχεια ερωτώνται εκ νέου για άλλους τέσσερις μήνες. Αυτό το σύστημα εναλλαγής επιτρέπει στο CPS να διατηρεί μια ισορροπία μεταξύ φρέσκων δεδομένων και διαχρονικής ανάλυσης, παρακολουθώντας τις αλλαγές με την πάροδο του χρόνου για συγκεκριμένα νοικοκυριά (Artola & Bell, 1999).

Η συλλογή δεδομένων στην CPS περιλαμβάνει λεπτομερείς συνεντεύξεις που διεξάγονται είτε αυτοπροσώπως είτε τηλεφωνικά. Οι συνεντευκτές ρωτούν για διάφορες πτυχές της δραστηριότητας στην αγορά εργασίας, όπως η κατάσταση απασχόλησης, οι ώρες εργασίας, οι δραστηριότητες αναζήτησης εργασίας και οι λόγοι ανεργίας. Η έρευνα συγκεντρώνει επίσης δημογραφικές πληροφορίες όπως η ηλικία, το φύλο, η φυλή, το μορφωτικό επίπεδο και η οικογενειακή κατάσταση. Αυτό το πλούσιο σύνολο δεδομένων επιτρέπει ολοκληρωμένες αναλύσεις των τάσεων της αγοράς εργασίας σε διάφορες ομάδες πληθυσμού (Chang & Krosnick, 2009).

Το μεγάλο μέγεθος του δείγματος της CPS είναι ένα από τα κύρια πλεονεκτήματά της, επιτρέποντας τη δημιουργία εύρωστων και αξιόπιστων στατιστικών για την αγορά εργασίας. Η εκτεταμένη κάλυψη διασφαλίζει ότι τα δεδομένα είναι αντιπροσωπευτικά για το σύνολο του πληθυσμού των ΗΠΑ, καθιστώντας την πολύτιμη πηγή για τους φορείς χάραξης πολιτικής και τους ερευνητές. Επιπλέον, ο διαχρονικός σχεδιασμός της CPS παρέχει πληροφορίες σχετικά με τη δυναμική της αγοράς εργασίας με την πάροδο του χρόνου, καταγράφοντας τάσεις και πρότυπα που μπορεί να διαφεύγουν από τις διατομικές έρευνες (Groves & Peytcheva, 2008).

Μία από τις κρίσιμες επιπτώσεις της ΚΥΠ είναι ο ρόλος της στην ενημέρωση της κυβερνητικής πολιτικής. Οι υπεύθυνοι χάραξης πολιτικής βασίζονται στα δεδομένα της CPS για να λάβουν τεκμηριωμένες αποφάσεις σχετικά με τις παρεμβάσεις στην αγορά εργασίας, τα προγράμματα κοινωνικής πρόνοιας και τις οικονομικές πολιτικές. Για παράδειγμα, τα ποσοστά ανεργίας που προκύπτουν από την CPS χρησιμοποιούνται για τον καθορισμό της επιλεξιμότητας για επιδόματα ανεργίας και για την κατανομή ομοσπονδιακών κονδυλίων στις πολιτείες. Τα δεδομένα βοηθούν επίσης στην αξιολόγηση της αποτελεσματικότητας των πολιτικών για την αγορά εργασίας και στον εντοπισμό τομέων που απαιτούν περαιτέρω παρέμβαση (Chang & Krosnick, 2009).

Χειροκίνητη συλλογή δεδομένων

Η χειροκίνητη συλλογή δεδομένων διαδραμάτισε ιστορικά σημαντικό ρόλο στην ανάλυση της αγοράς εργασίας, ιδίως μέσω συνεντεύξεων σε βάθος με εργοδότες. Αυτή η ποιοτική προσέγγιση περιλαμβάνει την άμεση συνεργασία με τους εργοδότες για τη συλλογή λεπτομερών πληροφοριών σχετικά με τις πρακτικές πρόσληψης, τις απαιτήσεις δεξιοτήτων και τις τάσεις απασχόλησης. Σε αντίθεση με τις ποσοτικές έρευνες, οι οποίες συχνά βασίζονται σε προκαθορισμένες ερωτήσεις και στατιστική δειγματοληψία, η χειροκίνητη συλλογή δεδομένων μέσω συνεντεύξεων επιτρέπει μια πιο λεπτή κατανόηση της δυναμικής της αγοράς εργασίας, αποτυπώνοντας τις πολυπλοκότητες και τις λεπτές αποχρώσεις που οι αριθμοί από μόνοι τους δεν μπορούν να αποδώσουν (Groves & Peytcheva, 2008).

Οι συνεντεύξεις σε βάθος με τους εργοδότες αποτελούν ακρογωνιαίο λίθο της παραδοσιακής έρευνας για την αγορά εργασίας, παρέχοντας πλούσιες, συμφραζόμενες πληροφορίες που είναι απαραίτητες για την ολοκληρωμένη κατανόηση των συνθηκών της αγοράς εργασίας. Οι συνεντεύξεις αυτές περιλαμβάνουν συνήθως δομημένες ή ημιδομημένες συζητήσεις με τους εργοδότες, κατά τη διάρκεια των οποίων οι ερευνητές ρωτούν για μια ποικιλία θεμάτων που σχετίζονται με τις πρακτικές απασχόλησης. Αυτά μπορεί να

περιλαμβάνουν τα είδη των δεξιοτήτων που έχουν ζήτηση, τις προκλήσεις που αντιμετωπίζουν κατά την πρόσληψη, την αποτελεσματικότητα των διαφόρων στρατηγικών πρόσληψης και τον αντίκτυπο των οικονομικών αλλαγών στις ανάγκες πρόσληψης (Artola & Bell, 1999).

Το βάθος των πληροφοριών που λαμβάνονται μέσω της χειροκίνητης συλλογής δεδομένων είναι ένα από τα κύρια πλεονεκτήματά της. Σε αντίθεση με τις ποσοτικές μεθόδους, οι οποίες μπορεί να περιορίζουν τις απαντήσεις σε σταθερές κατηγορίες, οι συνεντεύξεις επιτρέπουν στους εργοδότες να αναπτύξουν τις εμπειρίες και τις προοπτικές τους. Αυτά τα ποιοτικά δεδομένα μπορούν να αποκαλύψουν τους υποκείμενους παράγοντες που καθορίζουν τις τάσεις της αγοράς εργασίας, όπως οι αλλαγές στις απαιτήσεις της βιομηχανίας, τα αναδυόμενα κενά δεξιοτήτων και ο αντίκτυπος των τεχνολογικών εξελίξεων στις ανάγκες του εργατικού δυναμικού. Για παράδειγμα, οι συνεντεύξεις που διεξάγονται κατά τη μετάβαση στις ψηφιακές τεχνολογίες μπορεί να αποκαλύψουν συγκεκριμένες προκλήσεις που αντιμετωπίζουν οι εργοδότες στην εξεύρεση εργαζομένων με γνώσεις στο νέο λογισμικό ή τα νέα συστήματα, παρέχοντας πληροφορίες που μπορούν να χρησιμοποιηθούν για την κατάρτιση στοχευμένων προγραμμάτων κατάρτισης (Groves & Peytcheva, 2008).

Διοικητικά δεδομένα

Η ανάλυση των αρχείων απασχόλησης από τα δεδομένα κοινωνικής ασφάλισης αποτελεί μια παραδοσιακή και ιδιαίτερα αποτελεσματική προσέγγιση για την ανάλυση της αγοράς εργασίας. Τα διοικητικά δεδομένα από κυβερνητικές πηγές, όπως τα αρχεία κοινωνικής ασφάλισης, προσφέρουν πληθώρα πληροφοριών που είναι τόσο ακριβείς όσο και ολοκληρωμένες. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για την εξέταση των τάσεων απασχόλησης, των εισοδηματικών προτύπων και των δημογραφικών στοιχείων του εργατικού δυναμικού, παρέχοντας ανεκτίμητες πληροφορίες για την κατάσταση της αγοράς εργασίας (Devereux, 2007).

Τα δεδομένα κοινωνικής ασφάλισης είναι ιδιαίτερα πλεονεκτικά για την ανάλυση της αγοράς εργασίας λόγω της ακρίβειας και της πληρότητάς τους. Τα αρχεία αυτά συλλέγονται στο πλαίσιο της διαχείρισης των παροχών κοινωνικής ασφάλισης, πράγμα που σημαίνει ότι περιλαμβάνουν ένα ευρύ φάσμα πληροφοριών που σχετίζονται με την απασχόληση, συμπεριλαμβανομένων των αποδοχών, του ιστορικού απασχόλησης και των δημογραφικών στοιχείων των εργαζομένων. Αυτό το επίπεδο λεπτομέρειας επιτρέπει στους ερευνητές να διεξάγουν εις βάθος αναλύσεις των αποτελεσμάτων της αγοράς εργασίας. Επιπλέον, ο διαχρονικός χαρακτήρας των δεδομένων κοινωνικής ασφάλισης επιτρέπει την ανάλυση τάσεων για μεγάλες χρονικές περιόδους, προσφέροντας μια δυναμική εικόνα του τρόπου με τον οποίο εξελίσσονται τα πρότυπα

απασχόλησης και τα επίπεδα εισοδήματος με την πάροδο του χρόνου (Artola & Bell, 1999).

Τα οφέλη από τη χρήση διοικητικών δεδομένων από αρχεία κοινωνικής ασφάλισης στην ανάλυση της αγοράς εργασίας είναι πολλαπλά. Ένα από τα κύρια πλεονεκτήματα είναι το υψηλό επίπεδο ακρίβειας που εμπεριέχουν τα εν λόγω αρχεία. Δεδομένου ότι οι εργοδότες υποχρεούνται να αναφέρουν πληροφορίες για τις αποδοχές και την απασχόληση για σκοπούς φορολογίας και παροχών, τα δεδομένα είναι γενικά αξιόπιστα και απαλλαγμένα από τις ανακρίβειες που μπορεί να ταλαιπωρούν τα δεδομένα της έρευνας που υποβάλλονται από τους ίδιους τους εργοδότες. Αυτή η αξιοπιστία καθιστά τα δεδομένα κοινωνικής ασφάλισης μια εξαιρετική βάση για ισχυρές μελέτες της αγοράς εργασίας (Groves & Peytcheva, 2008).

Ένα άλλο σημαντικό πλεονέκτημα είναι ο ολοκληρωμένος χαρακτήρας των δεδομένων. Τα αρχεία κοινωνικής ασφάλισης καλύπτουν ένα ευρύ τμήμα του πληθυσμού, παρέχοντας ένα αντιπροσωπευτικό δείγμα που περιλαμβάνει διάφορα δημογραφικά στοιχεία, κλάδους και περιοχές. Αυτή η περιεκτικότητα επιτρέπει μια πιο ολιστική κατανόηση της αγοράς εργασίας, καταγράφοντας τις εμπειρίες διαφορετικών ομάδων που ενδέχεται να υποεκπροσωπούνται σε άλλες πηγές δεδομένων. Για παράδειγμα, οι αναλύσεις που χρησιμοποιούν δεδομένα κοινωνικής ασφάλισης μπορούν να αποκαλύψουν ανισότητες στην απασχόληση και το εισόδημα μεταξύ διαφορετικών φυλετικών και εθνοτικών ομάδων, φύλων και ηλικιακών κοορτών, αναδεικνύοντας τομείς στους οποίους μπορεί να απαιτούνται παρεμβάσεις πολιτικής.

Ο ρόλος του BLS στη διαμόρφωση της εργατικής πολιτικής των ΗΠΑ μέσω της συλλογής δεδομένων

Το Γραφείο Στατιστικών Εργασίας (BLS) έχει συμβάλει καθοριστικά στη διαμόρφωση της πολιτικής εργασίας των ΗΠΑ μέσω της ολοκληρωμένης και συστηματικής συλλογής δεδομένων για την αγορά εργασίας. Ως ο κύριος οργανισμός διερεύνησης στοιχείων για την ομοσπονδιακή κυβέρνηση στον τομέα των οικονομικών και στατιστικών της εργασίας, το BLS παράγει ένα ευρύ φάσμα εκθέσεων και συνόλων δεδομένων που χρησιμεύουν ως κρίσιμοι πόροι για τους φορείς χάραξης πολιτικής, τους ερευνητές και τις επιχειρήσεις. Αυτά τα δεδομένα έχουν βαθύτατο αντίκτυπο στη διαμόρφωση πολιτικών για την εργασία, συμπεριλαμβανομένων των αποφάσεων σχετικά με τους νόμους για τον κατώτατο μισθό, τα επιδόματα ανεργίας και τα προγράμματα ανάπτυξης του εργατικού δυναμικού (Artola & Bell, 1999).

Μια από τις σημαντικότερες συνεισφορές του BLS είναι το πρόγραμμα Στατιστικές Επαγγελματικής Απασχόλησης (OES). Το OES παρέχει λεπτομερή στοιχεία για την απασχόληση και τους μισθούς σε διάφορα επαγγέλματα και

κλάδους. Συλλέγοντας αυτά τα δεδομένα ετησίως μέσω μιας έρευνας σε εργοδότες, το BLS προσφέρει ένα ολοκληρωμένο στιγμιότυπο της αγοράς εργασίας των ΗΠΑ. Τα δεδομένα του OES είναι απαραίτητα για την κατανόηση της κατανομής των θέσεων εργασίας και των μισθών, τον εντοπισμό επαγγελματών υψηλής ζήτησης και την αξιολόγηση των μισθολογικών ανισοτήτων σε διάφορες περιοχές και δημογραφικές ομάδες. Οι υπεύθυνοι χάραξης πολιτικής χρησιμοποιούν αυτές τις πληροφορίες για τη χάραξη πολιτικών που αντιμετωπίζουν τις ανισοροπίες της αγοράς εργασίας, προωθούν τη δίκαιη αύξηση των μισθών και στοχεύουν τα προγράμματα κατάρτισης του εργατικού δυναμικού στις περιοχές με τις μεγαλύτερες ανάγκες (Groves & Peytcheva, 2008).

Ο αντίκτυπος των στοιχείων του BLS εκτείνεται πέρα από το OES. Το πρόγραμμα Current Employment Statistics (CES), μια άλλη βασική πρωτοβουλία, παρέχει μηνιαίες εκτιμήσεις της απασχόλησης, των ωρών εργασίας και των αποδοχών για το έθνος, τις πολιτείες και τις μητροπολιτικές περιοχές. Το πρόγραμμα αυτό παρέχει έγκαιρη πληροφόρηση για τις οικονομικές τάσεις και τις συνθήκες της αγοράς εργασίας, επιτρέποντας την ταχεία αντίδραση της πολιτικής σε αναδυόμενα ζητήματα. Για παράδειγμα, κατά τη διάρκεια της οικονομικής ύφεσης, τα στοιχεία του CES βοηθούν στον εντοπισμό των τομέων που πλήττονται περισσότερο από τις απώλειες θέσεων εργασίας, καθοδηγώντας τη χορήγηση επιδομάτων ανεργίας και άλλων μορφών οικονομικής ανακούφισης (Chang & Krosnick, 2009).

2.1.2 Περιορισμοί των παραδοσιακών μεθόδων

Κατανάλωση χρόνου

Οι παραδοσιακές μέθοδοι ανάλυσης των δεδομένων της αγοράς εργασίας, αν και θεμελιώδεις, αντιμετωπίζουν αρκετούς σημαντικούς περιορισμούς που μπορούν να επηρεάσουν την επικαιρότητα, την ακρίβεια και τη συνολική αξιοπιστία των πληροφοριών που παράγουν. Αυτές οι μέθοδοι, που συχνά περιλαμβάνουν χειροκίνητη συλλογή και επεξεργασία δεδομένων, πάσχουν από εγγενείς καθυστερήσεις και πιθανές προκαταλήψεις που μπορούν να στρεβλώσουν τα αποτελέσματα, οδηγώντας σε λιγότερο αντιπροσωπευτικά δεδομένα (Groves & Peytcheva, 2008).

Ένας από τους κύριους περιορισμούς των παραδοσιακών μεθόδων συλλογής δεδομένων για την αγορά εργασίας είναι η χρονοβόρα διαδικασία που απαιτείται. Οι εθνικές έρευνες, όπως αυτές που πραγματοποιούνται από το Γραφείο Στατιστικής Εργασίας (BLS), απαιτούν εκτεταμένο σχεδιασμό, συντονισμό και εκτέλεση. Η διαδικασία περιλαμβάνει συνήθως το σχεδιασμό της έρευνας, την

επιλογή αντιπροσωπευτικού δείγματος, την εκπαίδευση των συνεντευκτών, τη συλλογή των απαντήσεων και την επεξεργασία των δεδομένων. Καθένα από αυτά τα βήματα είναι χρονοβόρο και απαιτητικό σε πόρους, οδηγώντας συχνά σε σημαντικές καθυστερήσεις μεταξύ της συλλογής δεδομένων και της δημοσίευσης. Για παράδειγμα, η έρευνα Current Population Survey (CPS) παρέχει κρίσιμες πληροφορίες για την αγορά εργασίας, αλλά ο χρόνος που απαιτείται για τη συγκέντρωση και την ανάλυση των δεδομένων μπορεί να οδηγήσει σε καθυστέρηση, καθιστώντας δύσκολη την ταχεία ανταπόκριση στις αναδυόμενες τάσεις της αγοράς εργασίας ή στους οικονομικούς κλυδωνισμούς (Artola & Bell, 1999).

Αυτές οι καθυστερήσεις στη δημοσίευση των δεδομένων μπορεί να έχουν κρίσιμες συνέπειες για τους υπεύθυνους χάραξης πολιτικής και τους ερευνητές που βασίζονται σε ενημερωμένες πληροφορίες για να λάβουν τεκμηριωμένες αποφάσεις. Κατά τη διάρκεια περιόδων ταχείας οικονομικής αλλαγής, όπως οι υφέσεις ή οι τεχνολογικές διαταραχές, η ανάγκη για έγκαιρα στοιχεία γίνεται ακόμη πιο έντονη. Ωστόσο, η εγγενής καθυστέρηση των παραδοσιακών μεθόδων συλλογής δεδομένων σημαίνει ότι, όταν τα δεδομένα είναι διαθέσιμα, μπορεί να μην αντικατοπτρίζουν πλέον με ακρίβεια την τρέχουσα κατάσταση της αγοράς εργασίας, περιορίζοντας έτσι τη χρησιμότητά τους για άμεσες παρεμβάσεις πολιτικής (Groves & Peytcheva, 2008).

Πιθανές προκαταλήψεις

Ένας άλλος σημαντικός περιορισμός των παραδοσιακών μεθόδων είναι η πιθανότητα μεροληψίας που εισάγεται από το σχεδιασμό της έρευνας ή την επιρροή του υπεύθυνου της συνέντευξης. Οι έρευνες σχεδιάζονται με ένα σύνολο προκαθορισμένων ερωτήσεων και επιλογών απάντησης, οι οποίες μερικές φορές μπορεί να μην καταγράφουν την πλήρη πολυπλοκότητα των εμπειριών ή των προοπτικών των ερωτηθέντων. Αυτός ο σχεδιασμός μπορεί να εισάγει μεροληψίες, ιδίως εάν οι ερωτήσεις δεν έχουν ελεγχθεί επαρκώς ως προς την πολιτισμική ευαισθησία ή τη συνάφεια σε διαφορετικές δημογραφικές ομάδες. Για παράδειγμα, οι ερωτήσεις σχετικά με την κατάσταση απασχόλησης ή τις δραστηριότητες αναζήτησης εργασίας μπορεί να ερμηνευθούν διαφορετικά από τους ερωτηθέντες με βάση το πολιτισμικό τους υπόβαθρο ή τις προσωπικές τους συνθήκες, οδηγώντας σε ασυνέπειες στα δεδομένα που συλλέγονται (Chang & Krosnick, 2009).

Η επιρροή του συνεντευκτή είναι μια άλλη πηγή πιθανής μεροληψίας. Ο τρόπος με τον οποίο ο υπεύθυνος της συνέντευξης παρουσιάζει τις ερωτήσεις ή αλληλεπιδρά με τους ερωτηθέντες μπορεί να επηρεάσει ακούσια τις απαντήσεις τους. Για παράδειγμα, ο τόνος της φωνής του συνεντευκτή, η γλώσσα του σώματος ή ακόμη και ο τρόπος με τον οποίο επαναδιατυπώνει μια ερώτηση για

λόγους σαφήνειας μπορεί να επηρεάσει τον τρόπο με τον οποίο απαντούν οι ερωτώμενοι. Αυτές οι ανεπαίσθητες αλληλεπιδράσεις μπορούν να εισάγουν μεροληψίες στις απαντήσεις, οι οποίες μπορούν να διαστρεβλώσουν τα δεδομένα και να οδηγήσουν σε ανακριβή συμπεράσματα σχετικά με τις συνθήκες της αγοράς εργασίας (Devereux, 2007).

Η μεροληψία της δειγματοληψίας αποτελεί ιδιαίτερο πρόβλημα στις παραδοσιακές μεθόδους έρευνας. Η διαδικασία επιλογής ενός αντιπροσωπευτικού δείγματος είναι ζωτικής σημασίας για να εξασφαλιστεί ότι τα αποτελέσματα της έρευνας αντικατοπτρίζουν με ακρίβεια τον ευρύτερο πληθυσμό. Ωστόσο, η επίτευξη πραγματικής αντιπροσωπευτικότητας αποτελεί πρόκληση. Ορισμένες ομάδες μπορεί να υποεκπροσωπούνται ή να υπερεκπροσωπούνται λόγω του χρησιμοποιούμενου πλαισίου δειγματοληψίας ή επειδή ορισμένα άτομα είναι πιο πιθανό να συμμετάσχουν σε έρευνες από άλλα (Groves & Peytcheva, 2008).

Ακρίβεια και πληρότητα των δεδομένων

Ένας από τους θεμελιώδεις περιορισμούς των παραδοσιακών μεθόδων συλλογής δεδομένων για την αγορά εργασίας είναι το ζήτημα της ακρίβειας και της πληρότητας των δεδομένων, ιδίως όταν βασίζονται σε αυτοαναφερόμενες πληροφορίες. Οι έρευνες και τα ερωτηματολόγια, που αποτελούν τη ραχοκοκαλιά των παραδοσιακών μεθόδων συλλογής δεδομένων, συχνά εξαρτώνται από τα άτομα που θυμούνται και αναφέρουν με ακρίβεια το ιστορικό απασχόλησης, τους τίτλους εργασίας και τις ημερομηνίες απασχόλησης. Ωστόσο, η ανθρώπινη μνήμη είναι ασταθής και οι ερωτώμενοι μπορεί να παρέχουν εσφαλμένες ή ελλιπείς πληροφορίες είτε ακούσια είτε σκόπιμα (Artola & Bell, 1999).

Για παράδειγμα, οι ερωτηθέντες μπορεί να μην θυμούνται σωστά τον ακριβή τίτλο εργασίας τους ή τις ημερομηνίες που κατείχαν ορισμένες θέσεις. Ένα άτομο που εργαζόταν ως "Senior Data Analyst" μπορεί να αναφέρει απλώς "Data Analyst" λόγω λάθους ανάκλησης ή έλλειψης κατανόησης των αποχρώσεων μεταξύ των διαφόρων τίτλων. Παρομοίως, οι ημερομηνίες απασχόλησης μπορεί να αναφέρονται ανακριβώς, με τα άτομα είτε να συντομεύουν είτε να επιμηκύνουν τη θητεία τους σε ορισμένες θέσεις εργασίας, συχνά λόγω κενών μνήμης. Αυτές οι ανακρίβειες μπορούν να διαστρεβλώσουν σημαντικά τα δεδομένα, οδηγώντας σε εσφαλμένες αναλύσεις και λανθασμένες πολιτικές αποφάσεις (Groves & Peytcheva, 2008).

Επιπλέον, οι ερωτηθέντες μπορεί να έχουν κίνητρο να παρουσιάσουν τον εαυτό τους σε πιο ευνοϊκό φως, οδηγώντας σε σκόπιμη ψευδή αναφορά. Αυτό το φαινόμενο, γνωστό ως προκατάληψη κοινωνικής επιθυμητότητας, μπορεί να οδηγήσει σε υπερβολή των θετικών χαρακτηριστικών, όπως οι εργασιακές ευθύνες ή τα επιτεύγματα, και σε υποτίμηση των αρνητικών πτυχών, όπως οι

περίοδοι ανεργίας ή μερικής απασχόλησης. Τέτοιες μεροληψίες θέτουν περαιτέρω σε κίνδυνο την αξιοπιστία των δεδομένων που συλλέγονται με τις παραδοσιακές μεθόδους (Chang & Krosnick, 2009).

Η πληρότητα των δεδομένων είναι ένα άλλο κρίσιμο ζήτημα. Οι παραδοσιακές έρευνες αντιμετωπίζουν συχνά περιορισμούς όσον αφορά την ισότιμη προσέγγιση όλων των τμημάτων του πληθυσμού. Ορισμένες ομάδες, όπως οι παροδικοί πληθυσμοί, οι εργαζόμενοι χωρίς χαρτιά ή τα άτομα που απασχολούνται σε άτυπες ή συναυλιακές εργασίες, ενδέχεται να υποεκπροσωπούνται στα δείγματα των ερευνών. Αυτή η υποαντιπροσώπηση οδηγεί σε ελλιπή δεδομένα που δεν καταγράφουν το πλήρες εύρος της δυναμικής της αγοράς εργασίας. Για παράδειγμα, η οικονομία των συναυλιών έχει αυξηθεί σημαντικά, ωστόσο οι παραδοσιακές έρευνες ενδέχεται να μην καταγράφουν επαρκώς τις συνεισφορές και τις προκλήσεις των εργαζομένων στις συναυλίες λόγω του άτυπου χαρακτήρα της εργασίας τους (Bound et al., 1994).

Επιπλέον, η δομή και ο σχεδιασμός των ερευνών μπορεί να συμβάλει σε ελλιπή δεδομένα. Οι προκαθορισμένες κατηγορίες απαντήσεων μπορεί να μην περιλαμβάνουν όλους τους πιθανούς τίτλους εργασίας ή τα σενάρια απασχόλησης, αναγκάζοντας τους ερωτηθέντες να επιλέξουν την πλησιέστερη αντιστοιχία, η οποία μπορεί να μην είναι απολύτως ακριβής. Οι ερωτήσεις ανοικτού τύπου, αν και πιο ευέλικτες, συχνά οδηγούν σε ποικίλες και ασυνεπείς απαντήσεις που είναι δύσκολο να αναλυθούν συστηματικά (Aletdinova & Bakaev, 2019).

Έλλειψη πληροφοριών σε πραγματικό χρόνο

Ένας άλλος σημαντικός περιορισμός των παραδοσιακών μεθόδων συλλογής δεδομένων για την αγορά εργασίας είναι η αδυναμία τους να παρέχουν πληροφορίες σε πραγματικό χρόνο. Οι παραδοσιακές μέθοδοι, όπως οι έρευνες μεγάλης κλίμακας και η συλλογή διοικητικών δεδομένων, περιλαμβάνουν συνήθως χρονοβόρες διαδικασίες συλλογής, επικύρωσης και ανάλυσης δεδομένων. Αυτή η καθυστέρηση μπορεί να είναι προβληματική σε μια ταχέως μεταβαλλόμενη αγορά εργασίας, όπου τα έγκαιρα δεδομένα είναι ζωτικής σημασίας για την αποτελεσματική πολιτική και τη λήψη αποφάσεων (Alsultanny, 2013).

Οι παραδοσιακές μέθοδοι συχνά αποτυγχάνουν να καταγράψουν τις ταχείες αλλαγές και τις αναδυόμενες τάσεις στην αγορά εργασίας, όπως αυτές που παρατηρούνται στην οικονομία της συναυλίας. Η gig economy, που χαρακτηρίζεται από τις βραχυπρόθεσμες συμβάσεις και την ανεξάρτητη εργασία, έχει αυξηθεί εκθετικά τα τελευταία χρόνια. Ωστόσο, λόγω της άτυπης και συχνά παροδικής φύσης της συναυλιακής εργασίας, οι παραδοσιακές μέθοδοι συλλογής δεδομένων δυσκολεύονται να συμβαδίσουν με τις εξελίξεις της. Μέχρι να

συλλεχθούν, να επεξεργαστούν και να δημοσιευτούν τα δεδομένα της έρευνας, οι πληροφορίες μπορεί να είναι ήδη ξεπερασμένες, χάνοντας κρίσιμες αλλαγές στα πρότυπα απασχόλησης και τις οικονομικές συνθήκες Giorgashvili & Bakaev, 2018).

Για παράδειγμα, κατά τη διάρκεια οικονομικών κλυδωνισμών ή ταχείας τεχνολογικής εξέλιξης, η αγορά εργασίας μπορεί να υποστεί σημαντικές μεταβολές σε σύντομο χρονικό διάστημα. Η πανδημία COVID-19 είναι ένα σχετικό παράδειγμα, όπου οι παραδοσιακές μέθοδοι συλλογής δεδομένων καθυστέρησαν να καταγράψουν την ταχεία στροφή προς την εξ αποστάσεως εργασία, την αύξηση της ανεργίας και τις επακόλουθες αλλαγές στα πρότυπα ανάκτησης θέσεων εργασίας. Η ανάγκη για δεδομένα σε πραγματικό χρόνο έγινε εμφανής, καθώς οι υπεύθυνοι χάραξης πολιτικής απαιτούσαν άμεσες γνώσεις για την εφαρμογή έγκαιρων παρεμβάσεων και μέτρων στήριξης (Devereux, 2007).

Επιπλέον, ο στατικός χαρακτήρας των παραδοσιακών ερευνών σημαίνει ότι συνήθως διεξάγονται σε σταθερά χρονικά διαστήματα, όπως τριμηνιαία ή ετήσια. Ενώ αυτή η περιοδικότητα επιτρέπει τη διαχρονική ανάλυση, δεν παρέχει τη συνεχή ροή δεδομένων που είναι απαραίτητη για την παρακολούθηση των διακυμάνσεων της αγοράς εργασίας σε πραγματικό χρόνο. Αντίθετα, οι σύγχρονες τεχνικές συλλογής δεδομένων, όπως η απόξεση ιστοσελίδων, η ανάλυση των μέσων κοινωνικής δικτύωσης και η ανάλυση μεγάλων δεδομένων, προσφέρουν τη δυνατότητα συλλογής και ανάλυσης δεδομένων σε πραγματικό χρόνο, παρέχοντας μια δυναμική και επικαιροποιημένη εικόνα της αγοράς εργασίας (Giorgashvili & Bakaev, 2018).

Η έλλειψη γνώσης σε πραγματικό χρόνο από τις παραδοσιακές μεθόδους περιορίζει την αποτελεσματικότητά τους στην ενημέρωση για πολιτικές που ανταποκρίνονται και προσαρμόζονται. Οι φορείς χάραξης πολιτικής και οι επιχειρήσεις χρειάζονται έγκαιρες πληροφορίες για να αντιμετωπίσουν τις άμεσες προκλήσεις της αγοράς εργασίας και να αξιοποιήσουν τις αναδυόμενες ευκαιρίες. Χωρίς δεδομένα σε πραγματικό χρόνο, υπάρχει ο κίνδυνος να βασιστούν σε ξεπερασμένες πληροφορίες, με αποτέλεσμα καθυστερημένες αντιδράσεις και μη βέλτιστη λήψη αποφάσεων (Alsultanny, 2013).

2.2 Εργασίες NLP στην ανάλυση της αγοράς εργασίας

2.2.1 Εξέλιξη της NLP στο HR

Πρώιμες αναζητήσεις με λέξεις-κλειδιά

Στις πρώτες ημέρες της αξιοποίησης της επεξεργασίας φυσικής γλώσσας (NLP) για την ανάλυση της αγοράς εργασίας, οι αναζητήσεις με λέξεις-κλειδιά ήταν η

κύρια μέθοδος που χρησιμοποιούνταν για τον έλεγχο βιογραφικών σημειωμάτων. Αυτές οι αρχικές απόπειρες αυτοματοποίησης της διαδικασίας πρόσληψης βασίζονταν σε απλές τεχνικές αντιστοίχισης λέξεων-κλειδιών για το φιλτράρισμα βιογραφικών με βάση την παρουσία συγκεκριμένων όρων που οι υπεύθυνοι πρόσληψης θεωρούσαν σχετικούς με την περιγραφή της θέσης εργασίας. Τα πρώτα συστήματα παρακολούθησης υποψηφίων (ATS) αποτελούσαν την επιτομή αυτής της προσέγγισης, εστιάζοντας στον εντοπισμό και την καταμέτρηση των ακριβών αντιστοιχιών των λέξεων-κλειδιών στα βιογραφικά σημειώματα (Tanberk et al., 2023).

Η απλότητα του φιλτραρίσματος με βάση τις λέξεις-κλειδιά το έκανε ελκυστική λύση κατά την ίδρυσή του. Οι υπεύθυνοι προσλήψεων και οι διευθυντές προσλήψεων εισήγαγαν έναν κατάλογο λέξεων-κλειδιών που ήταν ζωτικής σημασίας για τη θέση εργασίας, όπως συγκεκριμένες δεξιότητες, πιστοποιήσεις ή τίτλους εργασίας. Στη συνέχεια, το ATS σάρωσε τα υποβληθέντα βιογραφικά για αυτές τις λέξεις-κλειδιά και κατέταξε τους υποψηφίους αναλόγως. Για παράδειγμα, εάν μια θέση εργασίας απαιτούσε επάρκεια στον προγραμματισμό Java, το σύστημα θα αναζητούσε τον όρο "Java" και θα έδινε προτεραιότητα στα βιογραφικά που τον περιείχαν. Η μέθοδος αυτή ήταν απλή και παρείχε έναν γρήγορο τρόπο για να κοσκινίζεται μεγάλος όγκος αιτήσεων, μειώνοντας σημαντικά τον χρόνο και την προσπάθεια που απαιτούνταν για χειροκίνητη εξέταση βιογραφικών σημειωμάτων (Deng et al., 2018).

Ωστόσο, η αποτελεσματικότητα των πρώτων αναζητήσεων με λέξεις-κλειδιά περιοριζόταν από διάφορους παράγοντες. Ένα σημαντικό ζήτημα ήταν η έλλειψη κατανόησης του πλαισίου. Τα συστήματα αντιστοίχισης λέξεων-κλειδιών δεν μπορούσαν να κάνουν διάκριση μεταξύ σχετικών και άσχετων εμφανίσεων μιας λέξης-κλειδιού. Για παράδειγμα, ένα βιογραφικό σημείωμα που ανέφερε τη λέξη "Java" στο πλαίσιο του "νησιού Java" θα αντιμετωπιζόταν το ίδιο με αυτό που ανέφερε τον "προγραμματισμό Java", οδηγώντας σε ψευδώς θετικά αποτελέσματα. Επιπλέον, αυτά τα συστήματα δεν ήταν σε θέση να αναγνωρίσουν συνώνυμα ή παραλλαγές των λέξεων-κλειδιών, πράγμα που σήμαινε ότι βιογραφικά σημειώματα που χρησιμοποιούσαν διαφορετικές ορολογίες μπορούσαν να αγνοηθούν. Για παράδειγμα, ένας υποψήφιος με εξειδίκευση στην "ανάπτυξη λογισμικού" μπορεί να μην αναγνωριζόταν εάν η λέξη-κλειδί που είχε καθοριστεί ήταν "μηχανική λογισμικού" (Tanberk et al., 2023).

Η ακαμψία των πρώτων αναζητήσεων με λέξεις-κλειδιά είχε συχνά ως αποτέλεσμα τον αποκλεισμό δυνητικά ισχυρών υποψηφίων των οποίων τα βιογραφικά δεν ταίριαζαν ακριβώς με τις προκαθορισμένες λέξεις-κλειδιά. Αυτός ο περιορισμός υπογράμμισε την ανάγκη για πιο εξελιγμένες μεθόδους που θα μπορούσαν να κατανοήσουν και να ερμηνεύσουν το πλαίσιο των λέξεων,

οδηγώντας στην ανάπτυξη πιο προηγμένων τεχνικών NLP στην ανάλυση της αγοράς εργασίας (Deng et al., 2018).

Εισαγωγή των στατιστικών μοντέλων

Για να αντιμετωπιστούν οι περιορισμοί της απλής αντιστοίχισης λέξεων-κλειδιών, η υιοθέτηση στατιστικών μοντέλων σηματοδότησε μια σημαντική πρόοδο στη χρήση του NLP για την ανάλυση της αγοράς εργασίας. Αυτά τα μοντέλα, όπως το Term Frequency-Inverse Document Frequency (TF-IDF) και το bag-of-words, εισήγαγαν μια πιο διαφοροποιημένη προσέγγιση στην ανάλυση κειμένου, λαμβάνοντας υπόψη τη συχνότητα και την κατανομή των λέξεων εντός ενός εγγράφου και σε ένα σώμα εγγράφων (Alghazal, 2021).

Το TF-IDF, για παράδειγμα, είναι ένα στατιστικό μέτρο που αξιολογεί τη σημασία μιας λέξης σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων (corpus). Αυτή η μέθοδος αποδίδει υψηλότερα βάρη σε λέξεις που εμφανίζονται συχνά σε ένα συγκεκριμένο έγγραφο αλλά είναι λιγότερο συχνές σε ολόκληρο το σώμα, προσδιορίζοντας έτσι όρους που είναι πιο αντιπροσωπευτικοί του μοναδικού περιεχομένου του εγγράφου. Στο πλαίσιο της εξέτασης βιογραφικών σημειωμάτων, η TF-IDF μπορεί να βοηθήσει στην ανάδειξη των σημαντικότερων δεξιοτήτων και εμπειριών ενός υποψηφίου, μειώνοντας την έμφαση στις συχνά χρησιμοποιούμενες λέξεις που είναι λιγότερο κατατοπιστικές. Για παράδειγμα, λέξεις όπως "υπεύθυνος" και "διαχειρίστηκε" μπορεί να εμφανίζονται σε πολλά βιογραφικά σημειώματα, αλλά το TF-IDF μπορεί να υποβαθμίσει αυτούς τους όρους, δίνοντας μεγαλύτερη σημασία σε εξειδικευμένες δεξιότητες όπως "μηχανική μάθηση" ή "ανάλυση δεδομένων" (Cambria & White, 2014).

Το μοντέλο bag-of-words ενίσχυσε περαιτέρω την ανάλυση κειμένου, αναπαριστώντας το κείμενο ως μια συλλογή λέξεων χωρίς να λαμβάνει υπόψη τη σειρά τους. Αυτό το μοντέλο δημιουργεί ένα διάνυσμα συχνότητας λέξεων για κάθε έγγραφο, επιτρέποντας την εύκολη σύγκριση και ανάλυση. Στις εφαρμογές ανθρώπινου δυναμικού, το μοντέλο bag-of-words διευκόλυνε την ανάπτυξη πιο εκλεπτυσμένων αλγορίθμων αντιστοίχισης που θα μπορούσαν να αποτυπώσουν καλύτερα το περιεχόμενο των βιογραφικών σημειωμάτων και των περιγραφών θέσεων εργασίας. Με τη μετατροπή του κειμένου σε αριθμητικά διανύσματα, το μοντέλο αυτό επέτρεψε τη χρήση διαφόρων τεχνικών στατιστικής και μηχανικής μάθησης για τον εντοπισμό μοτίβων και συσχετίσεων μεταξύ των προφίλ των υποψηφίων και των απαιτήσεων των θέσεων εργασίας (Ricca et al., 2004).

Για παράδειγμα, σε μια μελέτη που αναλύει αγγελίες θέσεων εργασίας και βιογραφικά σημειώματα, το μοντέλο "σακούλα λέξεων" θα μπορούσε να χρησιμοποιηθεί για τη δημιουργία προφίλ τόσο των περιγραφών θέσεων εργασίας όσο και των βιογραφικών σημειωμάτων των υποψηφίων. Συγκρίνοντας αυτά τα προφίλ, το σύστημα θα μπορούσε να αξιολογήσει την

ομοιότητα μεταξύ των προσόντων ενός υποψηφίου και των απαιτήσεων της θέσης εργασίας, παρέχοντας μια πιο ακριβή αντιστοίχιση από την απλή αναζήτηση με λέξεις-κλειδιά. Η προσέγγιση αυτή επέτρεψε επίσης τον εντοπισμό κοινών δεξιοτήτων και προσόντων σε διαφορετικούς ρόλους εργασίας, βοηθώντας στον προγραμματισμό και την ανάπτυξη του εργατικού δυναμικού (Ricca et al., 2004).

Η εισαγωγή στατιστικών μοντέλων, όπως το TF-IDF και το bag-of-words, αποτέλεσε καίρια μετατόπιση από την καθαρά συντακτική αντιστοίχιση λέξεων-κλειδιών στην πιο σημασιολογική κατανόηση του κειμένου. Αυτά τα μοντέλα βελτίωσαν την ακρίβεια και την ανάκληση των διαδικασιών διαλογής βιογραφικών σημειωμάτων λαμβάνοντας υπόψη το πλαίσιο και τη συνάφεια των λέξεων. Επιπλέον, έθεσαν τις βάσεις για την ανάπτυξη ακόμη πιο προηγμένων τεχνικών NLP, όπως οι αλγόριθμοι μηχανικής μάθησης και τα μοντέλα βαθιάς μάθησης, τα οποία συνεχίζουν να βελτιώνουν την αποτελεσματικότητα και την ακρίβεια της ανάλυσης της αγοράς εργασίας σήμερα (Alghazal, 2021).

Προώθηση σε μοντέλα μηχανικής μάθησης

Η μετάβαση από τις πρώτες αναζητήσεις λέξεων-κλειδιών και τα βασικά στατιστικά μοντέλα στη μηχανική μάθηση (ML) σηματοδότησε ένα σημαντικό άλμα στην πολυπλοκότητα της ανάλυσης κειμένου σε εφαρμογές της αγοράς εργασίας. Τα μοντέλα μηχανικής μάθησης έφεραν ένα νέο επίπεδο ακρίβειας και προσαρμοστικότητας σε εργασίες όπως ο έλεγχος βιογραφικών σημειωμάτων και η αντιστοίχιση θέσεων εργασίας, ξεπερνώντας τους περιορισμούς της αντιστοίχισης λέξεων-κλειδιών και των προσεγγίσεων με βάση τη συχνότητα. Αυτή η αλλαγή επέτρεψε τη βαθύτερη κατανόηση του πλαισίου και του νοήματος μέσα στα βιογραφικά σημειώματα και τις περιγραφές θέσεων εργασίας, βελτιώνοντας έτσι την ακρίβεια και τη συνάφεια των αντιστοιχίσεων μεταξύ υποψηφίων και θέσεων εργασίας (Alghazal, 2021).

Οι μηχανές διανυσμάτων υποστήριξης (SVM) και η λογιστική παλινδρόμηση είναι δύο από τις κύριες τεχνικές μηχανικής μάθησης που έχουν υιοθετηθεί ευρέως για εργασίες ταξινόμησης κειμένου σε βιογραφικά σημειώματα. Οι SVMs είναι ιδιαίτερα κατάλληλες για χώρους υψηλών διαστάσεων, γεγονός που τις καθιστά αποτελεσματικές για το χειρισμό της τεράστιας σειράς χαρακτηριστικών που παράγονται από δεδομένα κειμένου. Στο πλαίσιο του ελέγχου βιογραφικών σημειωμάτων, οι SVM μπορούν να ταξινομήσουν βιογραφικά σημειώματα μετατρέποντας το περιεχόμενο κειμένου σε διανύσματα χαρακτηριστικών και βρίσκοντας το υπερεπίπεδο που διαχωρίζει καλύτερα τις διάφορες κατηγορίες θέσεων εργασίας. Αυτό επιτρέπει στο μοντέλο να διακρίνει μεταξύ βιογραφικών σημειωμάτων για διάφορους ρόλους εργασίας, όπως η ανάπτυξη λογισμικού

έναντι της διαχείρισης έργων, με βάση το περιεχόμενο και τη δομή του κειμένου (Tanberk et al., 2023).

Η λογιστική παλινδρόμηση, μια άλλη ισχυρή τεχνική, έχει χρησιμοποιηθεί για τη δυαδική και πολυταξική ταξινόμηση κειμένων. Εκτιμά την πιθανότητα ότι μια δεδομένη είσοδος ανήκει σε μια συγκεκριμένη κατηγορία. Για παράδειγμα, η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί για να προβλεφθεί αν ένα βιογραφικό σημείωμα ταιριάζει με μια περιγραφή θέσης εργασίας ή για να ταξινομηθεί το επίπεδο εξειδίκευσης που διαθέτει ένας υποψήφιος σε έναν συγκεκριμένο τομέα δεξιοτήτων. Με την εκπαίδευση του μοντέλου σε δεδομένα με ετικέτες, η λογιστική παλινδρόμηση μπορεί να μάθει τις σχέσεις μεταξύ χαρακτηριστικών κειμένου και κατηγοριών θέσεων εργασίας, επιτρέποντας ακριβέστερο και αυτοματοποιημένο έλεγχο βιογραφικών σημειωμάτων (Alghazal, 2021).

Η εφαρμογή αυτών των μοντέλων μηχανικής μάθησης στο ανθρώπινο δυναμικό έχει επιδείξει σημαντικές βελτιώσεις σε σχέση με τις παραδοσιακές μεθόδους. Για παράδειγμα, μια μελέτη θα μπορούσε να δείξει πώς τα SVM υπερτερούν των συστημάτων που βασίζονται σε λέξεις-κλειδιά στον εντοπισμό σχετικών υποψηφίων για τεχνικές θέσεις, λαμβάνοντας υπόψη τις αποχρώσεις στον τρόπο με τον οποίο περιγράφονται οι δεξιότητες και οι εμπειρίες. Ομοίως, τα μοντέλα λογιστικής παλινδρόμησης μπορούν να εκπαιδευτούν ώστε να αναγνωρίζουν τη σημασία ορισμένων προσόντων και εμπειριών, παρέχοντας έτσι μια πιο διαφοροποιημένη αξιολόγηση της καταλληλότητας των υποψηφίων (Tanberk et al., 2023).

Τρέχοντα μοντέλα της τρέχουσας κατάστασης

Η συνεχής εξέλιξη του NLP έχει οδηγήσει στην ανάπτυξη ακόμη πιο προηγμένων μοντέλων, τα οποία αξιοποιούν τεχνικές βαθιάς μάθησης για την επίτευξη κορυφαίων επιδόσεων σε εργασίες ανάλυσης κειμένου. Τα συνεπτυγμένα νευρωνικά δίκτυα (CNNs) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs), συμπεριλαμβανομένων των παραλλαγών τους, όπως τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTMs), έχουν ιδιαίτερη απήχηση στις εφαρμογές HR. Τα CNN, που αρχικά σχεδιάστηκαν για την επεξεργασία εικόνας, έχουν προσαρμοστεί για εργασίες ταξινόμησης κειμένου. Είναι σε θέση να συλλάβουν τοπικά μοτίβα στο κείμενο μέσω επιπέδων συνελίξεων που ανιχνεύουν n-grams ή ομάδες λέξεων, οι οποίες είναι απαραίτητες για την κατανόηση του πλαισίου. Για παράδειγμα, ένα CNN μπορεί να αναγνωρίσει τη φράση "εμπειρία στη διαχείριση έργων" ως σημαντικό χαρακτηριστικό σε ένα βιογραφικό σημείωμα. Με τη στοίβαξη πολλαπλών στρωμάτων συνελίξεων, τα CNN μπορούν να μαθαίνουν όλο και πιο σύνθετες αναπαραστάσεις κειμένου, οδηγώντας σε ακριβέστερη

ταξινόμηση βιογραφικών σημειωμάτων στις κατάλληλες κατηγορίες θέσεων εργασίας (Deng et al., 2018).

Τα RNNs, και ιδιαίτερα τα LSTMs, υπερέχουν στο χειρισμό διαδοχικών δεδομένων, γεγονός που τα καθιστά ιδανικά για την επεξεργασία και κατανόηση βιογραφικών σημειωμάτων, τα οποία είναι εγγενώς δομημένα ως ακολουθίες λέξεων ή προτάσεων. Οι LSTM αντιμετωπίζουν τους περιορισμούς των παραδοσιακών RNNs, καταγράφοντας αποτελεσματικά τις μακροχρόνιες εξαρτήσεις στο κείμενο. Αυτό τους επιτρέπει να κατανοούν το πλαίσιο της εμπειρίας ενός υποψηφίου σε πολλαπλούς ρόλους εργασίας και να αναγνωρίζουν μοτίβα που καλύπτουν ολόκληρα βιογραφικά σημειώματα. Για παράδειγμα, οι LSTM μπορούν να διακρίνουν μεταξύ ενός υποψηφίου που έχει προοδεύσει σταθερά στην καριέρα του και ενός υποψηφίου που έχει αλλάξει συχνά θέσεις εργασίας, παρέχοντας βαθύτερες πληροφορίες για την επαγγελματική του πορεία (Alghazal, 2021).

Η ενσωμάτωση των CNN και των RNN οδήγησε στην ανάπτυξη υβριδικών μοντέλων που αξιοποιούν τα πλεονεκτήματα και των δύο αρχιτεκτονικών. Αυτά τα μοντέλα μπορούν να συλλάβουν τόσο τοπικά όσο και παγκόσμια μοτίβα στο κείμενο, ενισχύοντας την ικανότητά τους να ταξινομήσουν με ακρίβεια βιογραφικά σημειώματα. Για παράδειγμα, ένα υβριδικό μοντέλο μπορεί να χρησιμοποιεί στρώματα CNN για την εξαγωγή χαρακτηριστικών από μεμονωμένες προτάσεις και στρώματα RNN για την κατανόηση της συνολικής ακολουθίας αυτών των προτάσεων, συνδυάζοντας αυτές τις γνώσεις για να προβεί σε μια ολοκληρωμένη αξιολόγηση των προσόντων ενός υποψηφίου (Tanberk et al., 2023).

Τα τρέχοντα σύγχρονα μοντέλα επωφελούνται επίσης από τις τεχνικές εκμάθησης μεταφοράς, όπου τα μοντέλα που έχουν προ-εκπαιδευτεί σε μεγάλα σύνολα δεδομένων προσαρμόζονται για συγκεκριμένες εργασίες. Μοντέλα όπως το BERT (Bidirectional Encoder Representations from Transformers) έφεραν επανάσταση στο NLP παρέχοντας ενσωμάτωση σε συμφραζόμενα που αποτυπώνουν το διαφοροποιημένο νόημα των λέξεων με βάση το περιεχόμενό τους σε μια πρόταση. Η λεπτομερής ρύθμιση του BERT σε δεδομένα βιογραφικών σημειωμάτων του επιτρέπει να κατανοεί τη συγκεκριμένη ορολογία και δομή που χρησιμοποιείται στα βιογραφικά σημειώματα, οδηγώντας σε εξαιρετικά ακριβή αναγνώριση και ταξινόμηση οντοτήτων (Deng et al., 2018).

2.2.2 Τρέχουσες εφαρμογές NLP

Predictive Analytics για τις τάσεις πρόσληψης

Στο πεδίο της ανάλυσης της αγοράς εργασίας, η προγνωστική ανάλυση έχει αναδειχθεί σε ισχυρό εργαλείο, αξιοποιώντας την επεξεργασία φυσικής γλώσσας

(NLP) για την πρόβλεψη μελλοντικών τάσεων προσλήψεων με βάση τις δημοσιεύσεις θέσεων εργασίας και τα δεδομένα της αγοράς. Αυτή η εφαρμογή της NLP περιλαμβάνει την ανάλυση τεράστιου όγκου πληροφοριών κειμένου από αγγελίες εργασίας, βιογραφικά σημειώματα και άλλες πηγές δεδομένων της αγοράς εργασίας για τον εντοπισμό μοτίβων και την πραγματοποίηση τεκμηριωμένων προβλέψεων σχετικά με τις μελλοντικές τάσεις απασχόλησης. Εργαλεία όπως η Burning Glass Technologies αποτελούν παράδειγμα της χρήσης της NLP για την προγνωστική ανάλυση, παρέχοντας πολύτιμες πληροφορίες που βοηθούν τις επιχειρήσεις και τους υπεύθυνους χάραξης πολιτικής να προβλέψουν τις αλλαγές στην αγορά εργασίας (Cammaraat & Squicciarini, 2021).

Η Burning Glass Technologies χρησιμοποιεί προηγμένες τεχνικές NLP για την επεξεργασία και ανάλυση εκατομμυρίων αγγελιών εργασίας από διάφορες πηγές. Εξάγοντας και κατηγοριοποιώντας πληροφορίες όπως οι απαιτούμενες δεξιότητες, οι τίτλοι εργασίας και οι κλάδοι, η Burning Glass μπορεί να εντοπίσει τις αναδυόμενες τάσεις στην αγορά εργασίας. Για παράδειγμα, η πλατφόρμα μπορεί να εντοπίσει μια αυξανόμενη ζήτηση για συγκεκριμένες τεχνικές δεξιότητες, όπως η μηχανική μάθηση ή η κυβερνοασφάλεια, υποδεικνύοντας τομείς όπου οι ευκαιρίες απασχόλησης είναι πιθανό να αυξηθούν. Αυτές οι πληροφορίες επιτρέπουν στα εκπαιδευτικά ιδρύματα να προσαρμόσουν τα προγράμματά τους ώστε να καλύψουν τη μελλοντική ζήτηση, συμβάλλοντας στη γεφύρωση των κενών δεξιοτήτων και στην προετοιμασία του εργατικού δυναμικού για τις επερχόμενες ανάγκες της αγοράς εργασίας (Xing, Cambria & Welsch, 2017).

Επιπλέον, η προγνωστική ανάλυση μπορεί να αναδείξει τις αλλαγές στα πρότυπα απασχόλησης, όπως η αύξηση της εξ αποστάσεως εργασίας ή η αυξανόμενη επικράτηση των ρόλων της gig economy. Αναλύοντας τις δημοσιεύσεις θέσεων εργασίας σε βάθος χρόνου, τα εργαλεία NLP μπορούν να εντοπίσουν τάσεις στις απαιτήσεις θέσεων εργασίας και στις συνθήκες εργασίας, παρέχοντας μια μελλοντική άποψη της αγοράς εργασίας. Για παράδειγμα, κατά τη διάρκεια της πανδημίας COVID-19, η Burning Glass μπόρεσε να παρακολουθήσει την αύξηση των δημοσιεύσεων θέσεων εργασίας εξ αποστάσεως, προσφέροντας κρίσιμες πληροφορίες σχετικά με τον τρόπο με τον οποίο η πανδημία αναδιαμόρφωσε το εργατικό δυναμικό. Αυτές οι προβλέψεις βοήθησαν τις επιχειρήσεις να προσαρμόσουν τις στρατηγικές πρόσληψής τους και επέτρεψαν στους υπεύθυνους χάραξης πολιτικής να σχεδιάσουν παρεμβάσεις για την υποστήριξη των εργαζομένων εξ αποστάσεως (Alsayed & Awad, 2023).

Η χρήση της προγνωστικής ανάλυσης στις τάσεις πρόσληψης επεκτείνεται επίσης στον προγραμματισμό και την ανάπτυξη του εργατικού δυναμικού. Οι εταιρείες μπορούν να αξιοποιήσουν αυτές τις γνώσεις για να προβλέψουν τις

ανάγκες προσλήψεων και να σχεδιάσουν ανάλογα. Για παράδειγμα, εάν η προγνωστική ανάλυση υποδεικνύει μελλοντική έλλειψη επιστημόνων δεδομένων, οι επιχειρήσεις μπορούν να επενδύσουν προληπτικά σε προγράμματα κατάρτισης ή να προσαρμόσουν τις προσπάθειές τους για προσλήψεις ώστε να προσελκύσουν ταλέντα σε αυτόν τον τομέα. Αυτή η προληπτική προσέγγιση όχι μόνο βοηθά τις εταιρείες να παραμείνουν ανταγωνιστικές, αλλά εξασφαλίζει επίσης ένα πιο ισορροπημένο και προετοιμασμένο εργατικό δυναμικό (Cammeraat & Squicciarini, 2021).

Περιγραφή εργασίας Ανάλυση

Η ανάλυση περιγραφών θέσεων εργασίας είναι μια άλλη κρίσιμη εφαρμογή της NLP στην ανάλυση της αγοράς εργασίας, η οποία επικεντρώνεται στην κατανόηση και τη βελτίωση του περιεχομένου των δημοσιεύσεων θέσεων εργασίας για την προσέλκυση καλύτερων υποψηφίων. Πλατφόρμες όπως η Textio χρησιμοποιούν NLP για να αναλύουν τις περιγραφές θέσεων εργασίας και να παρέχουν πληροφορίες σχετικά με τον τρόπο βελτίωσής τους, καθιστώντας τις πιο ελκυστικές για ένα ευρύ φάσμα υποψηφίων και αυξάνοντας την πιθανότητα προσέλκυσης κορυφαίων ταλέντων (Shaikh et al., 2021).

Η Textio χρησιμοποιεί προηγμένους αλγόριθμους NLP για να αξιολογεί τις περιγραφές θέσεων εργασίας ως προς διάφορους παράγοντες, όπως η αναγνωσιμότητα, ο τόνος και η περιεκτικότητα. Αναλύοντας τη γλώσσα που χρησιμοποιείται στις προκηρύξεις θέσεων εργασίας, η Textio μπορεί να εντοπίσει όρους και φράσεις που μπορεί να αποτρέπουν ακούσια ορισμένες ομάδες υποψηφίων. Για παράδειγμα, η υπερβολικά τεχνική ορολογία ή η γλώσσα με βάση το φύλο μπορεί να αποθαρρύνει τους υποψήφιους με προσόντα από το να υποβάλουν αίτηση. Το Textio παρέχει σε πραγματικό χρόνο προτάσεις για την αναδιατύπωση ή την αντικατάσταση τέτοιων όρων, βοηθώντας τις εταιρείες να δημιουργήσουν περιγραφές θέσεων εργασίας που είναι πιο περιεκτικές και ελκυστικές σε ένα ευρύτερο κοινό (Valdez-Almada et al., 2017).

Η ανάλυση της πλατφόρμας υπερβαίνει την απλή αντιστοίχιση λέξεων-κλειδιών, αξιοποιώντας μοντέλα βαθιάς μάθησης για την κατανόηση του πλαισίου και του συναισθήματος του κειμένου. Αυτό επιτρέπει στο Textio να προσφέρει διαφοροποιημένες συστάσεις που ενισχύουν τη συνολική αποτελεσματικότητα των περιγραφών θέσεων εργασίας. Για παράδειγμα, εάν μια δημοσίευση θέσης εργασίας δίνει έμφαση στις πολλές ώρες εργασίας ή σε περιβάλλοντα υψηλής πίεσης, η Textio μπορεί να προτείνει τρόπους εξισορρόπησης αυτών των πληροφοριών με θετικές πτυχές της εταιρικής κουλτούρας ή ευκαιρίες ανάπτυξης, καθιστώντας την δημοσίευση πιο ελκυστική για τους δυνητικούς υποψηφίους (Tanberk et al., 2023).

Η ανάλυση της περιγραφής θέσεων εργασίας βοηθά επίσης στον εντοπισμό των βασικών δεξιοτήτων και προσόντων που απαιτούνται για έναν ρόλο. Εξάγοντας και κατηγοριοποιώντας αυτά τα στοιχεία, τα εργαλεία NLP μπορούν να διασφαλίσουν ότι οι προκηρύξεις θέσεων εργασίας μεταφέρουν με σαφήνεια και ακρίβεια τις βασικές απαιτήσεις, μειώνοντας την πιθανότητα υποβολής αιτήσεων που δεν ταιριάζουν. Αυτή η ακρίβεια στις περιγραφές θέσεων εργασίας όχι μόνο βελτιώνει την ποιότητα των αιτήσεων, αλλά και εξορθολογίζει τη διαδικασία πρόσληψης, εξοικονομώντας χρόνο και πόρους για τις ομάδες ανθρώπινου δυναμικού (Shaikh et al., 2021).

Επιπλέον, πλατφόρμες όπως η Textio μπορούν να συγκρίνουν τις περιγραφές θέσεων εργασίας με τα πρότυπα του κλάδου, παρέχοντας πληροφορίες σχετικά με το πώς οι δημοσιεύσεις μιας εταιρείας συγκρίνονται με εκείνες των ανταγωνιστών. Αυτή η ανταγωνιστική ανάλυση βοηθά τις επιχειρήσεις να κατανοήσουν πού βρίσκονται στην αγορά εργασίας και να κάνουν τις απαραίτητες προσαρμογές για να προσελκύσουν κορυφαία ταλέντα. Για παράδειγμα, εάν διαπιστωθεί ότι οι περιγραφές θέσεων εργασίας μιας εταιρείας είναι λιγότερο ελκυστικές ή περιεκτικές από εκείνες των κορυφαίων ανταγωνιστών, η επιχείρηση μπορεί να εφαρμόσει τις συστάσεις της Textio για να βελτιώσει τις δημοσιεύσεις της και να ενισχύσει την ελκυστικότητά τους για τους υποψηφίους (Valdez-Almada et al., 2017).

Ανάλυση βιογραφικού σημειώματος

Η ανάλυση βιογραφικών σημειωμάτων είναι μια κρίσιμη εφαρμογή της Επεξεργασίας Φυσικής Γλώσσας (NLP) στην ανάλυση της αγοράς εργασίας, όπου χρησιμοποιούνται προηγμένες τεχνικές για την εξαγωγή δομημένων δεδομένων από βιογραφικά σημειώματα. Η διαδικασία αυτή περιλαμβάνει τη μετατροπή αδόμητου κειμένου βιογραφικού σημειώματος σε δομημένη μορφή που μπορεί εύκολα να αναλυθεί και να υποστεί επεξεργασία από αυτοματοποιημένα συστήματα. Εργαλεία όπως το Sonren και το Rchilli έχουν καταστεί ηγέτες του κλάδου σε αυτόν τον τομέα, προσφέροντας εξελιγμένες λύσεις λογισμικού που βελτιώνουν την αποτελεσματικότητα και την ακρίβεια της ανάλυσης βιογραφικών σημειωμάτων (Tanberk et al., 2023).

Οι Sonren και Rchilli χρησιμοποιούν αλγόριθμους NLP για να αναλύουν βιογραφικά σημειώματα και να εντοπίζουν βασικά στοιχεία όπως προσωπικές πληροφορίες, εκπαίδευση, εργασιακή εμπειρία, δεξιότητες και πιστοποιήσεις. Αναλύοντας αυτά τα δεδομένα, τα εργαλεία αυτά δημιουργούν δομημένα προφίλ που μπορούν εύκολα να ενσωματωθούν σε συστήματα παρακολούθησης αιτούντων (ATS) και άλλες πλατφόρμες ανθρώπινου δυναμικού. Αυτά τα δομημένα δεδομένα επιτρέπουν την αποτελεσματικότερη και ακριβέστερη διαλογή των υποψηφίων, επιτρέποντας στους υπεύθυνους προσλήψεων να

εντοπίζουν γρήγορα τους πιο κατάλληλους υποψηφίους για μια συγκεκριμένη θέση εργασίας (Pendyala et al., 2022).

Η διαδικασία της ανάλυσης βιογραφικού σημειώματος αρχίζει με βήματα προεπεξεργασίας, όπου το κείμενο καθαρίζεται και κανονικοποιείται για να εξασφαλιστεί η συνοχή. Αυτό περιλαμβάνει την αφαίρεση άσχετων πληροφοριών, τη διόρθωση κοινών προβλημάτων μορφοποίησης και την τυποποίηση όρων. Για παράδειγμα, οι διαφορετικοί τρόποι αναγραφής ενός τίτλου εργασίας, όπως "Μηχανικός λογισμικού" και "Προγραμματιστής λογισμικού", κανονικοποιούνται σε μια τυποποιημένη μορφή. Στη συνέχεια, οι αλγόριθμοι NLP αναλύουν το κείμενο, εντοπίζοντας και κατηγοριοποιώντας τις σχετικές οντότητες. Η αναγνώριση ονομαστικών οντοτήτων (NER) παίζει καθοριστικό ρόλο εδώ, καθώς βοηθά στην αναγνώριση συγκεκριμένων οντοτήτων, όπως ονόματα, ημερομηνίες και τεχνικές δεξιότητες (Anusha et al., 2023).

Ένα από τα σημαντικά πλεονεκτήματα της χρήσης εργαλείων όπως το Sonren και το Rchilli είναι η ικανότητά τους να χειρίζονται βιογραφικά σε διάφορες μορφές, όπως PDF, Word και απλό κείμενο. Αυτή η ευελιξία διασφαλίζει ότι τα δεδομένα από όλους τους τύπους βιογραφικών μπορούν να εξαχθούν και να δομηθούν με ακρίβεια. Επιπλέον, αυτά τα εργαλεία έχουν σχεδιαστεί για να κατανοούν και να επεξεργάζονται βιογραφικά από διαφορετικούς κλάδους και περιοχές, φιλοξενώντας ένα ευρύ φάσμα ορολογιών και μορφών (Tanberk et al., 2023).

Αγορές ταλέντων και μηχανές αντιστοίχισης

Μια άλλη μετασχηματιστική εφαρμογή της NLP στην ανάλυση της αγοράς εργασίας είναι οι αγορές ταλέντων και οι μηχανές αντιστοίχισης. Πλατφόρμες όπως το LinkedIn χρησιμοποιούν προηγμένους αλγόριθμους NLP για την αντιστοίχιση υποψηφίων με ευκαιρίες απασχόλησης, ενισχύοντας την αποδοτικότητα και την αποτελεσματικότητα της διαδικασίας πρόσληψης. Η μηχανή συστάσεων του LinkedIn, για παράδειγμα, αξιοποιεί τον NLP για να αναλύει τα προφίλ των χρηστών, τις δημοσιεύσεις θέσεων εργασίας και διάφορες άλλες πηγές δεδομένων για να παρέχει εξατομικευμένες συστάσεις θέσεων εργασίας και αντιστοιχίσεις υποψηφίων (Bharathi, 2022).

Η μηχανή συστάσεων του LinkedIn χρησιμοποιεί εξελιγμένα μοντέλα NLP για την κατανόηση του περιεχομένου και του πλαισίου των προφίλ των χρηστών και των περιγραφών εργασίας. Αυτό περιλαμβάνει την ανάλυση της εργασιακής εμπειρίας ενός χρήστη, των δεξιοτήτων του, των εγκρίσεων και άλλων στοιχείων του προφίλ του, ώστε να δημιουργηθεί μια ολοκληρωμένη εικόνα των επαγγελματικών του προσόντων. Ομοίως, οι αγγελίες θέσεων εργασίας αναλύονται για να εξαχθούν οι βασικές απαιτήσεις, οι αρμοδιότητες και τα προσόντα. Συγκρίνοντας αυτές τις δομημένες αναπαραστάσεις, η μηχανή

συστάσεων μπορεί να εντοπίσει τις καλύτερες αντιστοιχίες μεταξύ υποψηφίων και ευκαιριών απασχόλησης (Gugnani & Misra, 2020).

Μία από τις βασικές τεχνολογίες που στηρίζουν τη μηχανή αντιστοίχισης του LinkedIn είναι η χρήση μοντέλων μετασχηματιστών, όπως το BERT (Bidirectional Encoder Representations from Transformers). Η ικανότητα του BERT να κατανοεί το πλαίσιο και τις σημασιολογικές σχέσεις μέσα στο κείμενο επιτρέπει την ακριβή αντιστοίχιση των υποψηφίων με τις σχετικές αγγελίες θέσεων εργασίας, ακόμη και όταν η ακριβής ορολογία που χρησιμοποιείται στα προφίλ και στις περιγραφές θέσεων εργασίας διαφέρει. Για παράδειγμα, η BERT μπορεί να αναγνωρίσει ότι οι λέξεις "ανάλυση δεδομένων" και "ανάλυση δεδομένων" αναφέρονται στο ίδιο σύνολο δεξιοτήτων, βελτιώνοντας την ακρίβεια των αντιστοιχίσεων (Skondras et al., 2023).

Ο αντίκτυπος των μηχανών αντιστοίχισης με βάση το NLP στις αγορές ταλέντων είναι βαθύς. Παρέχοντας εξατομικευμένες συστάσεις θέσεων εργασίας, οι πλατφόρμες αυτές ενισχύουν τη δέσμευση και την ικανοποίηση των χρηστών. Οι αναζητούντες εργασία έχουν περισσότερες πιθανότητες να βρουν ευκαιρίες που ταιριάζουν με τις δεξιότητες και τις επαγγελματικές τους φιλοδοξίες, ενώ οι εργοδότες λαμβάνουν αιτήσεις από πιο κατάλληλους υποψηφίους. Αυτή η βελτιωμένη διαδικασία αντιστοίχισης όχι μόνο εξορθολογίζει την πρόσληψη, αλλά αυξάνει επίσης την πιθανότητα επιτυχούς τοποθέτησης σε θέσεις εργασίας, προς όφελος τόσο των αναζητούντων εργασία όσο και των εργοδοτών (Bharathi, 2022).

Επιπλέον, τα δεδομένα που παράγονται από αυτές τις πλατφόρμες μπορούν να χρησιμοποιηθούν για τον εντοπισμό ευρύτερων τάσεων και γνώσεων της αγοράς εργασίας. Για παράδειγμα, το Economic Graph του LinkedIn αξιοποιεί δεδομένα από τη μηχανή αντιστοίχισης για να παρέχει σε πραγματικό χρόνο πληροφορίες σχετικά με τη δυναμική της παγκόσμιας αγοράς εργασίας, όπως αναδυόμενες δεξιότητες, ζητούμενους ρόλους εργασίας και περιφερειακές τάσεις απασχόλησης. Αυτές οι πληροφορίες είναι πολύτιμες για τους υπεύθυνους χάραξης πολιτικής, τους εκπαιδευτικούς και τις επιχειρήσεις, βοηθώντας τους να κατανοήσουν και να ανταποκριθούν στις μεταβαλλόμενες συνθήκες της αγοράς εργασίας (Gugnani & Misra, 2020).

2.3 Εργασία αναγνώρισης ονομαστικών οντοτήτων (NER)

2.3.1 Βασικές αρχές της NER

Σκοπός της NER

Η αναγνώριση ονομαστικών οντοτήτων (NER) είναι μια θεμελιώδης εργασία στην Επεξεργασία Φυσικής Γλώσσας (NLP) που περιλαμβάνει τον εντοπισμό και την ταξινόμηση ονομαστικών οντοτήτων σε κείμενο σε προκαθορισμένες κατηγορίες, όπως ονόματα προσώπων, οργανισμοί, τοποθεσίες, ημερομηνίες και άλλα. Ο πρωταρχικός σκοπός της NER είναι η εξαγωγή δομημένων πληροφοριών από μη δομημένο κείμενο, επιτρέποντας τη μετατροπή των ακατέργαστων δεδομένων κειμένου σε μορφή χρήσιμη για διάφορες μεταγενέστερες εφαρμογές. Με τη μετατροπή του κειμένου σε δομημένες οντότητες, η NER διευκολύνει την αποτελεσματικότερη ανάκτηση πληροφοριών, την ανάλυση δεδομένων και τις αυτοματοποιημένες διαδικασίες λήψης αποφάσεων (Su et al., 2022).

Η NER διαδραματίζει κρίσιμο ρόλο σε πολλές εφαρμογές του πραγματικού κόσμου. Για παράδειγμα, στο πλαίσιο των βιογραφικών σημειωμάτων και των περιγραφών θέσεων εργασίας, η NER μπορεί να χρησιμοποιηθεί για τον εντοπισμό βασικών οντοτήτων, όπως τίτλοι θέσεων εργασίας, ονόματα εταιρειών και συγκεκριμένες δεξιότητες. Αυτά τα δομημένα δεδομένα μπορούν στη συνέχεια να χρησιμοποιηθούν για την ακριβέστερη αντιστοίχιση των υποψηφίων με τις κενές θέσεις εργασίας, ενισχύοντας την αποτελεσματικότητα της διαδικασίας πρόσληψης. Ομοίως, στον τομέα των οικονομικών, η NER μπορεί να βοηθήσει στην εξαγωγή σημαντικών πληροφοριών από οικονομικές εκθέσεις, άρθρα ειδήσεων και κανονιστικές καταθέσεις, παρέχοντας πληροφορίες για τις τάσεις της αγοράς και την οργανωτική απόδοση (Abdallah et al., 2012).

Μια άλλη σημαντική εφαρμογή της NER είναι στον τομέα της ανάκτησης πληροφοριών και των μηχανών αναζήτησης. Αναγνωρίζοντας οντότητες μέσα σε ερωτήματα αναζήτησης και έγγραφα, οι μηχανές αναζήτησης μπορούν να παρέχουν πιο συναφή αποτελέσματα στους χρήστες. Για παράδειγμα, εάν ένας χρήστης αναζητήσει "εστιατόρια στη Νέα Υόρκη", ένα σύστημα NER μπορεί να αναγνωρίσει τα "εστιατόρια" ως τύπο οντότητας και τη "Νέα Υόρκη" ως τοποθεσία, βελτιώνοντας έτσι τα αποτελέσματα αναζήτησης ώστε να περιλαμβάνουν σχετικές επιχειρήσεις εστίασης στην καθορισμένη περιοχή (Yu et al., 2022).

Στον τομέα της υγειονομικής περίθαλψης, η NER χρησιμοποιείται για την εξαγωγή κρίσιμων πληροφοριών από ιατρικούς φακέλους, ερευνητικά έγγραφα και εκθέσεις κλινικών δοκιμών. Οντότητες όπως ονόματα ασθενειών, φάρμακα και πληροφορίες ασθενών αναγνωρίζονται και ταξινομούνται, επιτρέποντας την καλύτερη διαχείριση και ανάλυση δεδομένων. Αυτά τα δομημένα δεδομένα μπορούν να βελτιώσουν τη φροντίδα των ασθενών, να υποστηρίξουν την κλινική έρευνα και να διευκολύνουν την ανάπτυξη νέων θεραπειών (Abdallah et al., 2012).

Η ΝΕΡ ενισχύει επίσης τις δυνατότητες των chatbots και των εικονικών βοηθών, επιτρέποντάς τους να κατανοούν και να απαντούν αποτελεσματικότερα στα ερωτήματα των χρηστών. Με τον εντοπισμό οντοτήτων στην είσοδο του χρήστη, τα συστήματα αυτά μπορούν να παρέχουν πιο ακριβείς και σχετικές με το πλαίσιο απαντήσεις. Για παράδειγμα, ένας εικονικός βοηθός μπορεί να αναγνωρίσει το αίτημα ενός χρήστη να "κλείσει μια πτήση για το Παρίσι" αναγνωρίζοντας την "πτήση" ως τύπο οντότητας που σχετίζεται με τα ταξίδια και το "Παρίσι" ως προορισμό, βελτιώνοντας έτσι τη διαδικασία κράτησης. (Su et al., 2022)

Κοινοί αλγόριθμοι

Η ανάπτυξη των συστημάτων ΝΕΡ έχει εξελιχθεί σημαντικά με την πάροδο των ετών, με διάφορες προσεγγίσεις να χρησιμοποιούνται για τη βελτίωση της ακρίβειας και της αποτελεσματικότητάς τους. Οι προσεγγίσεις αυτές μπορούν να κατηγοριοποιηθούν σε γενικές γραμμές σε μεθόδους βασισμένες σε κανόνες, στατιστικές μεθόδους και μεθόδους μηχανικής μάθησης, καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς (Nie et al., 2021).

Οι προσεγγίσεις που βασίζονται σε κανόνες ήταν από τις πρώτες μεθόδους που χρησιμοποιήθηκαν για τη ΝΕΡ. Τα συστήματα αυτά βασίζονται σε προκαθορισμένους κανόνες και πρότυπα για τον εντοπισμό και την ταξινόμηση οντοτήτων στο κείμενο. Για παράδειγμα, ένα σύστημα βασισμένο σε κανόνες μπορεί να χρησιμοποιεί κανονικές εκφράσεις για την αναγνώριση προτύπων, όπως λέξεις με κεφαλαία για ονόματα προσώπων ή συγκεκριμένες λέξεις-κλειδιά για οργανισμούς. Ενώ τα συστήματα που βασίζονται σε κανόνες μπορεί να είναι αποτελεσματικά για συγκεκριμένες εργασίες με σαφώς καθορισμένα πρότυπα, συχνά δυσκολεύονται με τη μεταβλητότητα και την πολυπλοκότητα της φυσικής γλώσσας. Είναι επίσης δύσκολο να κλιμακωθούν και να συντηρηθούν, καθώς οι κανόνες πρέπει να διαμορφώνονται και να ενημερώνονται χειροκίνητα (Pirouani & Oliveira, 2021).

Τα στατιστικά μοντέλα εισήγαγαν μια πιο ευέλικτη προσέγγιση στην ΕΠΑ, αξιοποιώντας πιθανοτικές μεθόδους για τον εντοπισμό οντοτήτων με βάση το περιεχόμενό τους μέσα στο κείμενο. Τα κρυφά μοντέλα Markov (HMM) είναι ένα αξιοσημείωτο παράδειγμα αυτής της προσέγγισης. Τα HMM χρησιμοποιούν πιθανότητες για να μοντελοποιήσουν την ακολουθία των λέξεων και τις αντίστοιχες ετικέτες οντοτήτων, επιτρέποντας στο σύστημα να καταγράφει τις εξαρτήσεις μεταξύ των λέξεων. Παρόλο που τα HMMs βελτίωσαν τα συστήματα που βασίζονται σε κανόνες χειριζόμενα ένα ευρύτερο φάσμα παραλλαγών στο κείμενο, εξακολουθούν να αντιμετωπίζουν περιορισμούς στη σύλληψη εξαρτήσεων μεγάλης εμβέλειας και σύνθετων μοτίβων (Li et al., 2018).

Η έλευση της μηχανικής μάθησης, ιδίως της μάθησης με επίβλεψη, σηματοδότησε σημαντική πρόοδο στην ΕΑΑ. Τα Conditional Random Fields (CRFs) έγιναν δημοφιλής επιλογή για εργασίες NER λόγω της ικανότητάς τους να μοντελοποιούν την υπό συνθήκη πιθανότητα των ακολουθιών ετικετών δεδομένου του κειμένου εισόδου. Τα CRF μπορούν να καταγράψουν πολύπλοκες εξαρτήσεις μεταξύ των λέξεων και των ετικετών τους, οδηγώντας σε ακριβέστερη αναγνώριση οντοτήτων. Με την εκπαίδευση σε σχολιασμένα σύνολα δεδομένων, οι CRF μαθαίνουν τα μοτίβα και τις σχέσεις που καθορίζουν τους διαφορετικούς τύπους οντοτήτων, με αποτέλεσμα την ισχυρή απόδοση σε διάφορους τομείς κειμένου (Nie et al., 2021).

Τα μοντέλα βαθιάς μάθησης έφεραν περαιτέρω επανάσταση στην ΕΑΑ αξιοποιώντας τα νευρωνικά δίκτυα για την αυτόματη εκμάθηση αναπαραστάσεων χαρακτηριστικών από το ακατέργαστο κείμενο. Τα δίκτυα αμφίδρομης μακράς βραχυπρόθεσμης μνήμης (BiLSTM) είναι ιδιαίτερα αποτελεσματικά για την ΕΑΑ λόγω της ικανότητάς τους να συλλαμβάνουν το πλαίσιο τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις. Τα BiLSTM επεξεργάζονται το κείμενο και προς τις δύο κατευθύνσεις, επιτρέποντας στο μοντέλο να κατανοήσει το πλήρες πλαίσιο κάθε λέξης, το οποίο είναι ζωτικής σημασίας για την ακριβή αναγνώριση οντοτήτων. Αυτά τα μοντέλα εκπαιδεύονται σε μεγάλα σχολιασμένα σώματα κειμένων, επιτρέποντάς τους να γενικεύουν καλά σε νέο, αθέατο κείμενο (Pirovani & Oliveira, 2021).

Η εισαγωγή μοντέλων μετασχηματιστών, όπως το BERT (Bidirectional Encoder Representations from Transformers), έχει ωθήσει τα όρια της NER ακόμη περισσότερο. Η βαθιά αμφίδρομη αρχιτεκτονική του BERT του επιτρέπει να καταγράφει περίπλοκες σχέσεις πλαισίου μέσα στο κείμενο, οδηγώντας σε κορυφαίες επιδόσεις σε εργασίες NER. Με την προ-εκπαίδευση σε τεράστιες ποσότητες δεδομένων κειμένου και τη λεπτομερή ρύθμιση σε συγκεκριμένα σύνολα δεδομένων NER, η BERT μπορεί να αναγνωρίζει και να ταξινομεί με ακρίβεια οντότητες σε διάφορους τομείς και γλώσσες (Nie et al., 2021).

Μετρικές αξιολόγησης

Η αξιολόγηση της απόδοσης των συστημάτων αναγνώρισης ονομαστικών οντοτήτων (NER) είναι ζωτικής σημασίας για την κατανόηση της αποτελεσματικότητας και της αξιοπιστίας τους. Για να επιτευχθεί αυτό, χρησιμοποιούνται διάφορες μετρικές αξιολόγησης, καθεμία από τις οποίες παρέχει μια διαφορετική οπτική γωνία για την ακρίβεια και την πληρότητα του συστήματος. Οι πρωταρχικές μετρικές που χρησιμοποιούνται στην αξιολόγηση των συστημάτων NER είναι η ακρίβεια, η ανάκληση και η βαθμολογία F1. Αυτές οι μετρικές προσφέρουν συλλογικά μια ισορροπημένη άποψη της ικανότητας του συστήματος να εντοπίζει και να ταξινομεί σωστά οντότητες μέσα σε ένα κείμενο.

Η ακρίβεια είναι το μέτρο της ακρίβειας των οντοτήτων που αναγνωρίζονται από το σύστημα NER. Υπολογίζεται ως ο λόγος των ορθά αναγνωρισμένων οντοτήτων προς τον συνολικό αριθμό των οντοτήτων που αναγνωρίστηκαν. Η υψηλή ακρίβεια δείχνει ότι οι οντότητες που αναγνωρίζονται από το σύστημα είναι ως επί το πλείστον σωστές, με λίγα ψευδώς θετικά αποτελέσματα. Για παράδειγμα, εάν ένα σύστημα NER αναγνωρίζει 100 οντότητες ως ονόματα προσώπων και 90 από αυτές είναι πράγματι ονόματα προσώπων, η ακρίβεια θα είναι 0,90 ή 90%. Η ακρίβεια είναι ιδιαίτερα σημαντική σε εφαρμογές όπου τα ψευδώς θετικά αποτελέσματα είναι δαπανηρά ή παραπλανητικά, όπως στην ανάλυση ιατρικών ή νομικών εγγράφων (Li et al., 2018).

Η ανάκληση, από την άλλη πλευρά, μετρά την ικανότητα του συστήματος να εντοπίζει όλες τις σχετικές οντότητες μέσα στο κείμενο. Είναι ο λόγος των σωστά αναγνωρισμένων οντοτήτων προς το συνολικό αριθμό των πραγματικών οντοτήτων που υπάρχουν. Η υψηλή ανάκληση υποδηλώνει ότι το σύστημα καταγράφει επιτυχώς τις περισσότερες σχετικές οντότητες, με λίγα ψευδώς αρνητικά αποτελέσματα. Για παράδειγμα, εάν υπάρχουν 100 ονόματα προσώπων σε ένα έγγραφο και το σύστημα NER αναγνωρίζει σωστά 80 από αυτά, η ανάκληση θα είναι 0,80 ή 80%. Η ανάκληση είναι κρίσιμη σε περιπτώσεις όπου η έλλειψη μιας οντότητας μπορεί να έχει σημαντικές συνέπειες, όπως στην ανάκτηση πληροφοριών ή στην εξόρυξη δεδομένων (Higashiyama et al., 2015).

Η βαθμολογία F1 είναι ένας αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, παρέχοντας μια ενιαία μετρική που εξισορροπεί και τις δύο πτυχές. Είναι ιδιαίτερα χρήσιμο όταν η ακρίβεια και η ανάκληση είναι εξίσου σημαντικές και πρέπει να εξετάζονται μαζί. Η βαθμολογία F1 υπολογίζεται ως εξής

$$F1=2 \times (\text{Ακρίβεια} + \text{Ανάκληση} \times \text{Ακρίβεια} \times \text{Ανάκληση}).$$

Αυτή η μετρική βοηθά στην αξιολόγηση της συνολικής αποτελεσματικότητας του συστήματος NER συνδυάζοντας την ακρίβεια και την ανάκληση σε ένα ενιαίο μέτρο. Για παράδειγμα, εάν ένα σύστημα NER έχει ακρίβεια 0,90 και ανάκληση 0,80, η βαθμολογία F1 θα είναι 0,85, υποδεικνύοντας μια καλά ισορροπημένη απόδοση (Raja et al., 2019).

Για να καταδείξετε την εφαρμογή αυτών των μετρικών, θεωρήστε την αξιολόγηση ενός μοντέλου NER που επεξεργάζεται ένα σώμα βιογραφικών σημειωμάτων. Ας υποθέσουμε ότι το σύστημα εντοπίζει 500 δεξιότητες, από τις οποίες οι 450 είναι σωστές (αληθή θετικά), οι 50 είναι λανθασμένες (ψευδώς θετικά) και του διαφεύγουν 100 σχετικές δεξιότητες (ψευδώς αρνητικά). Η ακρίβεια θα είναι 450/500, ή 90%, ενώ η ανάκληση θα είναι 450/550, ή περίπου 82%. Η βαθμολογία F1, που υπολογίζεται με τη χρήση αυτών των τιμών, θα παρείχε μια ολοκληρωμένη εικόνα της απόδοσης του συστήματος, εξισορροπώντας την ανάγκη να αναγνωριστούν σωστά όσο το δυνατόν

περισσότερες σχετικές δεξιότητες και ελαχιστοποιώντας ταυτόχρονα τον αριθμό των εσφαλμένων αναγνώρισεων (Manjula & Shridhar, 2019).

2.3.2 NER στο πλαίσιο του HR

Ανάλυση βιογραφικού σημειώματος

Στο πλαίσιο του ανθρώπινου δυναμικού (HR), η αναγνώριση ονομαστικών οντοτήτων (NER) διαδραματίζει καθοριστικό ρόλο στην αυτοματοποίηση και τη βελτίωση της διαδικασίας ανάλυσης βιογραφικών σημειωμάτων. Η ανάλυση βιογραφικού σημειώματος περιλαμβάνει την εξαγωγή σχετικών πληροφοριών από βιογραφικά σημειώματα, όπως δεξιότητες, τίτλους θέσεων εργασίας και οργανωτικές σχέσεις, και τη μετατροπή αυτών των μη δομημένων δεδομένων σε δομημένη μορφή που μπορεί εύκολα να αναλυθεί και να υποστεί επεξεργασία από συστήματα ανθρώπινου δυναμικού. Με την αξιοποίηση της NER, οι οργανισμοί μπορούν να εκσυγχρονίσουν τις διαδικασίες πρόσληψης, να μειώσουν τη χειρωνακτική προσπάθεια και να βελτιώσουν την ακρίβεια των αξιολογήσεων των υποψηφίων (Zhao et al., 2015).

Τα συστήματα NER είναι ικανά να εντοπίζουν συγκεκριμένες οντότητες μέσα σε βιογραφικά σημειώματα, όπως δεξιότητες, προσόντα, τίτλους θέσεων εργασίας και ονόματα εταιρειών. Για παράδειγμα, ένα βιογραφικό σημείωμα μπορεί να αναφέρει την "ανάλυση δεδομένων" ως δεξιότητα ή τον "μηχανικό λογισμικού" ως τίτλο εργασίας. Ένας αναλυτής βιογραφικού σημειώματος με δυνατότητα NER μπορεί να αναγνωρίσει με ακρίβεια αυτούς τους όρους και να τους κατηγοριοποιήσει κατάλληλα. Αυτή η ικανότητα είναι ζωτικής σημασίας για τη δημιουργία λεπτομερών προφίλ υποψηφίων που αναδεικνύουν τα δυνατά σημεία και τις επαγγελματικές εμπειρίες ενός ατόμου. Με την αυτοματοποίηση αυτής της εργασίας, οι οργανισμοί μπορούν να διασφαλίσουν ότι δεν παραλείπονται κρίσιμες πληροφορίες, καθιστώντας έτσι τη διαδικασία διαλογής πιο ολοκληρωμένη και αποτελεσματική (Li et al., 2018).

Η εφαρμογή της NER στην ανάλυση βιογραφικού σημειώματος περιλαμβάνει συνήθως διάφορα βήματα. Πρώτον, το κείμενο υφίσταται προεπεξεργασία για την τυποποίηση των μορφοτύπων και την αφαίρεση άσχετων πληροφοριών. Αυτή η προεπεξεργασία μπορεί να περιλαμβάνει τη μετατροπή των εγγράφων σε μια συνεπή μορφή, τη διόρθωση τυπογραφικών λαθών και την ομαλοποίηση των παραλλαγών στην ορολογία. Αφού γίνει η προεπεξεργασία του κειμένου, το σύστημα NER εφαρμόζει τα εκπαιδευμένα μοντέλα του για τον εντοπισμό και την ταξινόμηση οντοτήτων. Για παράδειγμα, ένα μοντέλο NER που έχει εκπαιδευτεί σε δεδομένα ειδικά για το ανθρώπινο δυναμικό μπορεί να αναγνωρίσει και να

εξάγει οντότητες που σχετίζονται με τεχνικές δεξιότητες, εκπαιδευτικά προσόντα και προηγούμενη απασχόληση (Sougandh et al., 2023).

Ένα πρακτικό παράδειγμα NER στην ανάλυση βιογραφικών σημειωμάτων μπορεί να παρατηρηθεί σε λογισμικό ανθρώπινου δυναμικού όπως το Sonren και το Rchilli. Αυτές οι πλατφόρμες χρησιμοποιούν προηγμένες τεχνικές NER για την ανάλυση βιογραφικών σημειωμάτων και την εξαγωγή δομημένων δεδομένων που συμπληρώνουν τα προφίλ των αιτούντων σε ένα σύστημα παρακολούθησης αιτούντων (ATS). Αυτά τα δομημένα δεδομένα επιτρέπουν στους υπεύθυνους προσλήψεων να πραγματοποιούν αποτελεσματικότερες αναζητήσεις και φίλτρα βάσει συγκεκριμένων κριτηρίων, όπως η εύρεση υποψηφίων με εμπειρία στη "διαχείριση έργων" ή με δεξιότητες "μηχανικής μάθησης". Η ακρίβεια και η αποτελεσματικότητα αυτών των συστημάτων βοηθούν τους οργανισμούς να εντοπίζουν γρήγορα τους κατάλληλους υποψηφίους από μια μεγάλη δεξαμενή υποψηφίων, επιταχύνοντας έτσι τη διαδικασία πρόσληψης (Boddu et al., 2023).

Αντιστοίχιση θέσεων εργασίας

Πέρα από την ανάλυση βιογραφικών σημειωμάτων, η NER παίζει σημαντικό ρόλο στη διαδικασία αντιστοίχισης θέσεων εργασίας, όπου χρησιμοποιείται για την ευθυγράμμιση των προφίλ των υποψηφίων με τις απαιτήσεις της θέσης εργασίας. Αυτό περιλαμβάνει τη σύγκριση των δομημένων δεδομένων που εξάγονται από τα βιογραφικά σημειώματα με τις πληροφορίες που περιέχονται στις περιγραφές θέσεων εργασίας για να προσδιοριστεί η καλύτερη δυνατή καταλληλότητα για μια δεδομένη θέση. Με τη χρήση του NER για την ανάλυση τόσο των βιογραφικών σημειωμάτων όσο και των αγγελιών θέσεων εργασίας, τα συστήματα ανθρώπινου δυναμικού μπορούν να εκτελούν ακριβέστερη και διαφοροποιημένη αντιστοίχιση, ενισχύοντας τη συνολική αποτελεσματικότητα των προσπαθειών πρόσληψης (Jiechieu & Tsopzé, 2020).

Τα συστήματα NER μπορούν να αναλύουν τις περιγραφές θέσεων εργασίας για να εντοπίζουν τις βασικές απαιτήσεις, όπως οι απαραίτητες δεξιότητες, οι τίτλοι θέσεων εργασίας και τα προσόντα. Για παράδειγμα, μια προκήρυξη θέσης εργασίας για έναν "μηχανικό λογισμικού" μπορεί να προσδιορίζει δεξιότητες όπως "προγραμματισμός σε Java", "ευέλικτη μεθοδολογία" και "ηγεσία ομάδας". Το σύστημα NER εξάγει αυτές τις οντότητες και τις αντιπαραβάλλει με τις οντότητες που προσδιορίζονται στα βιογραφικά των υποψηφίων. Η διαδικασία αυτή βοηθά στη δημιουργία ενός σύντομου καταλόγου υποψηφίων των οποίων τα προφίλ ευθυγραμμίζονται στενά με τις απαιτήσεις της θέσης εργασίας, βελτιώνοντας έτσι τη συνάφεια των αντιστοιχίσεων (Manjula & Shridhar, 2019).

Ένα ενδεικτικό παράδειγμα της αντιστοίχισης θέσεων εργασίας με τη χρήση NER είναι η μηχανή συστάσεων του LinkedIn. Το LinkedIn χρησιμοποιεί προηγμένες τεχνικές NER για την ανάλυση των προφίλ των χρηστών και των δημοσιεύσεων

θέσεων εργασίας, εξάγοντας οντότητες όπως δεξιότητες, τίτλους θέσεων εργασίας και ονόματα εταιρειών. Στη συνέχεια, η πλατφόρμα χρησιμοποιεί αυτά τα δομημένα δεδομένα για να προτείνει θέσεις εργασίας στους χρήστες και να προτείνει υποψηφίους στους υπεύθυνους προσλήψεων. Για παράδειγμα, εάν μια θέση εργασίας απαιτεί δεξιότητες "ανάλυσης δεδομένων" και εμπειρία με την "SQL", το σύστημα NER του LinkedIn εντοπίζει αυτές τις απαιτήσεις και τις αντιστοιχίζει με προφίλ υποψηφίων που έχουν αναφέρει αυτές τις δεξιότητες στα βιογραφικά τους. Αυτό έχει ως αποτέλεσμα εξαιρετικά σχετικές συστάσεις θέσεων εργασίας και αντιστοιχίσεις υποψηφίων, ενισχύοντας την αποτελεσματικότητα της διαδικασίας πρόσληψης (Jagwani et al., 2023).

Οργανωτική ανάλυση

Στο πλαίσιο του ανθρώπινου δυναμικού (HR), η αναγνώριση ονομαστικών οντοτήτων (NER) είναι ένα ανεκτίμητο εργαλείο για την οργανωτική ανάλυση. Η εφαρμογή αυτή περιλαμβάνει τον εντοπισμό και την εξαγωγή βασικών οντοτήτων που σχετίζονται με εταιρείες και ρόλους από βιογραφικά σημειώματα, γεγονός που βοηθά στη δημιουργία ολοκληρωμένων προφίλ υποψηφίων. Η ικανότητα ακριβούς αναγνώρισης οργανωτικών ονομάτων και ρόλων εργασίας όχι μόνο ενισχύει την αποτελεσματικότητα της διαδικασίας πρόσληψης, αλλά παρέχει επίσης βαθύτερες γνώσεις σχετικά με τις πορείες καριέρας και τα επαγγελματικά δίκτυα των δυνητικών προσλήψεων (Zhao et al., 2015).

Ένα από τα πρωταρχικά καθήκοντα στην οργανωτική ανάλυση είναι η εξαγωγή των ονομάτων των εταιρειών στις οποίες οι υποψήφιοι έχουν εργαστεί στο παρελθόν. Οι πληροφορίες αυτές είναι ζωτικής σημασίας για την αξιολόγηση της εμπειρίας του υποψηφίου και την κατανόηση του πλαισίου των προηγούμενων ρόλων του. Για παράδειγμα, η εργασία σε μια φημισμένη εταιρεία τεχνολογίας όπως η Google ή η Microsoft μπορεί να σηματοδοτεί υψηλό επίπεδο τεχνογνωσίας και επάρκειας, δεδομένων των αυστηρών προτύπων πρόσληψης και του ανταγωνιστικού περιβάλλοντος αυτών των εταιρειών. Τα συστήματα NER μπορούν να αναγνωρίζουν αυτόματα αυτά τα οργανωτικά ονόματα από τα βιογραφικά σημειώματα, διασφαλίζοντας ότι αυτές οι κρίσιμες πληροφορίες καταγράφονται και δομούνται με ακρίβεια (Sougandh et al., 2023).

Εκτός από την εξαγωγή ονομάτων εταιρειών, η NER χρησιμοποιείται για τον εντοπισμό ρόλων και τίτλων εργασίας σε βιογραφικά σημειώματα. Οι τίτλοι θέσεων εργασίας παρέχουν πολύτιμες πληροφορίες σχετικά με τις αρμοδιότητες και τα επίπεδα αρχαιότητας που κατείχαν οι υποψήφιοι στις προηγούμενες θέσεις τους. Για παράδειγμα, η διάκριση μεταξύ ενός "μηχανικού λογισμικού" και ενός "ανώτερου μηχανικού λογισμικού" είναι σημαντική, καθώς υποδηλώνει διαφορετικά επίπεδα εμπειρίας και ευθύνης. Με την ανάλυση των βιογραφικών σημειωμάτων για την εξαγωγή αυτών των τίτλων θέσεων εργασίας, τα

συστήματα NER βοηθούν στη δημιουργία λεπτομερών προφίλ υποψηφίων που αντικατοπτρίζουν όχι μόνο τους οργανισμούς για τους οποίους έχουν εργαστεί αλλά και τους συγκεκριμένους ρόλους που κατείχαν (Suhās & Manjunath, 2020). Η εξαγωγή οργανωτικών ονομάτων και ρόλων εργασίας διευκολύνει επίσης τη δημιουργία δομημένων βάσεων δεδομένων που μπορούν να χρησιμοποιηθούν για διάφορες λειτουργίες ανθρώπινου δυναμικού (Tran et al., 2022). Για παράδειγμα, μια πλατφόρμα προσλήψεων μπορεί να χρησιμοποιήσει αυτά τα δομημένα δεδομένα για να κατηγοριοποιήσει τους υποψηφίους με βάση την εμπειρία τους στον κλάδο ή τα επίπεδα αρχαιότητας. Αυτή η κατηγοριοποίηση επιτρέπει στους υπεύθυνους προσλήψεων να φιλτράρουν και να αναζητούν γρήγορα υποψηφίους που πληρούν συγκεκριμένα κριτήρια, όπως εκείνους με διευθυντική εμπειρία στον κλάδο της υγειονομικής περίθαλψης. Αυτοματοποιώντας την εξαγωγή και τη δόμηση αυτών των πληροφοριών, τα συστήματα NER μειώνουν σημαντικά τη χειροκίνητη προσπάθεια που απαιτείται για αυτές τις εργασίες, οδηγώντας σε ταχύτερες και ακριβέστερες αξιολογήσεις υποψηφίων (Jiechieu & Tsorzé, 2020). Ένα ενδεικτικό παράδειγμα της NEA στην οργανωτική ανάλυση μπορεί να παρατηρηθεί σε προηγμένες πλατφόρμες προσλήψεων όπως το LinkedIn. Το LinkedIn χρησιμοποιεί NER για την ανάλυση των προφίλ των χρηστών, εξάγοντας οντότητες όπως ονόματα εταιρειών και τίτλους θέσεων εργασίας για τη δημιουργία λεπτομερών επαγγελματικών προφίλ. Αυτά τα δομημένα δεδομένα χρησιμοποιούνται στη συνέχεια για την τροφοδοσία των αλγορίθμων σύστασης του LinkedIn, οι οποίοι προτείνουν ευκαιρίες απασχόλησης στους χρήστες και πιθανούς υποψηφίους στους υπεύθυνους προσλήψεων. Για παράδειγμα, εάν ένας υπεύθυνος προσλήψεων αναζητά υποψηφίους με εμπειρία σε κορυφαία χρηματοπιστωτικά ιδρύματα, το σύστημα NER του LinkedIn μπορεί να εντοπίσει και να επισημάνει προφίλ χρηστών που έχουν εργαστεί σε εταιρείες όπως η Goldman Sachs ή η JPMorgan Chase (Zhao et al., 2015).

2.4 Μεγάλα γλωσσικά μοντέλα (LLM)

2.4.1 Εισαγωγή στα LLM

Τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ) αντιπροσωπεύουν μια σημαντική πρόοδο στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (ΕΠΦ), που χαρακτηρίζεται από την ικανότητά τους να παράγουν κείμενο που μοιάζει με ανθρώπινο κείμενο και να εκτελούν μια ποικιλία σύνθετων εργασιών που σχετίζονται με τη γλώσσα. Τα μοντέλα αυτά βασίζονται σε αρχιτεκτονικές βαθιάς μάθησης, κυρίως μετασχηματιστές, που τους επιτρέπουν να επεξεργάζονται και να παράγουν κείμενο με αξιοσημείωτη ευχέρεια και συνοχή. Μεταξύ των πιο αξιοσημείωτων

LLM είναι τα GPT-3, BERT και T5, καθένα από τα οποία επιδεικνύει μοναδικές δυνατότητες και εφαρμογές στο NLP (Wild et al., 2021).

Τα LLM όπως το GPT-3, που αναπτύχθηκε από την OpenAI, αποτελούν παράδειγμα της δυνατότητας αυτών των μοντέλων να κατανοούν και να παράγουν κείμενο που είναι σχετικό με το πλαίσιο και γλωσσικά εξελιγμένο. Το GPT-3 βασίζεται σε μια αρχιτεκτονική μετασχηματιστή, που περιλαμβάνει 175 δισεκατομμύρια παραμέτρους, καθιστώντας το ένα από τα μεγαλύτερα και ισχυρότερα γλωσσικά μοντέλα μέχρι σήμερα. Αυτό το εκτεταμένο δίκτυο επιτρέπει στο GPT-3 να παράγει συνεκτικές παραγράφους κειμένου, να εκτελεί εργασίες όπως η μετάφραση, η περίληψη και ακόμη και η δημιουργική γραφή με ελάχιστες υποδείξεις. Για παράδειγμα, το GPT-3 μπορεί να παράγει δοκίμια, να απαντά σε ερωτήσεις και να δημιουργεί συνομιλιακούς πράκτορες που συμμετέχουν σε ρεαλιστικούς διαλόγους, επιδεικνύοντας την ευελιξία του σε διάφορες εφαρμογές NLP (Jagwani et al., 2023).

Η διαδικασία εκπαίδευσης των LLM περιλαμβάνει την τροφοδότηση του μοντέλου με τεράστιες ποσότητες δεδομένων κειμένου, επιτρέποντάς του να μάθει τις αποχρώσεις της γλώσσας, της γραμματικής και των συμφραζομένων. Αυτή η διαδικασία, γνωστή ως προ-εκπαίδευση, εφοδιάζει το μοντέλο με μια ευρεία κατανόηση της γλώσσας (Yao et al., 2022). Από την άλλη πλευρά, η λεπτομερής προσαρμογή περιλαμβάνει την προσαρμογή του προ-εκπαιδευμένου μοντέλου σε συγκεκριμένες εργασίες εκπαιδευοντάς το σε ένα στενότερο σύνολο δεδομένων σχετικών με αυτές τις εργασίες. Για παράδειγμα, το μοντέλο BERT (Bidirectional Encoder Representations from Transformers) μπορεί να ρυθμιστεί λεπτομερώς για συγκεκριμένες εφαρμογές, όπως η απάντηση ερωτήσεων ή η ταξινόμηση κειμένου. Κατά τη διάρκεια της τελειοποίησης, ο BERT προσαρμόζει τις παραμέτρους του ώστε να αποδίδει καλά στα δεδομένα που αφορούν συγκεκριμένες εργασίες, βελτιώνοντας την ακρίβεια και την αποδοτικότητά του (Gao et al., 2023).

Η αρχιτεκτονική του BERT διαφέρει από εκείνη του GPT-3 σε σημαντικούς τρόπους. Ενώ η GPT-3 είναι ένα μονόδρομο μοντέλο, που επεξεργάζεται κείμενο από αριστερά προς τα δεξιά, η BERT είναι αμφίδρομη, δηλαδή επεξεργάζεται κείμενο και προς τις δύο κατευθύνσεις ταυτόχρονα. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο BERT να αποκτήσει βαθύτερη κατανόηση του πλαισίου κάθε λέξης μέσα σε μια πρόταση, καθιστώντας το ιδιαίτερα αποτελεσματικό για εργασίες που απαιτούν λεπτή κατανόηση, όπως η ανάλυση συναισθήματος και η αναγνώριση οντοτήτων. Η τελειοποίηση του BERT για την απάντηση ερωτήσεων περιλαμβάνει την εκπαίδευσή του σε σύνολα δεδομένων, όπου το μοντέλο μαθαίνει να παρέχει ακριβείς και κατάλληλες απόψεις

πλασίου απαντήσεις σε ερωτήσεις, αποδεικνύοντας την προσαρμοστικότητά του σε διάφορες προκλήσεις NLP (Konstantinidis et al., 2022).

Η σύγκριση διαφορετικών LLM, όπως η GPT-3, η BERT και η T5, αποκαλύπτει διαφορετικές αρχιτεκτονικές επιλογές και προβλεπόμενες περιπτώσεις χρήσης. Η GPT-3, με την τεράστια κλίμακα και τις παραγωγικές της δυνατότητες, υπερέχει σε εργασίες που απαιτούν παραγωγή κειμένου και δημιουργικότητα. Η ικανότητά του να παράγει μεγάλα αποσπάσματα συνεκτικού κειμένου το καθιστά κατάλληλο για εφαρμογές στη δημιουργία περιεχομένου, τα chatbots και τη γλωσσική μετάφραση. Αντίθετα, η δύναμη του BERT έγκειται στην κατανόηση του πλαισίου εντός των προτάσεων, καθιστώντας το ιδανικό για εργασίες που απαιτούν βαθιά κατανόηση, όπως η ανάλυση συναισθήματος, η ταξινόμηση κειμένου και η απάντηση ερωτήσεων (Valmeekam et al., 2022).

Ο T5 (Text-To-Text Transfer Transformer), που αναπτύχθηκε από την Google, ακολουθεί μια διαφορετική προσέγγιση, διαμορφώνοντας όλες τις εργασίες NLP ως προβλήματα μετατροπής κειμένου σε κείμενο. Αυτό σημαίνει ότι τόσο η είσοδος όσο και η έξοδος αντιμετωπίζονται ως συμβολοσειρές κειμένου, επιτρέποντας στον T5 να ρυθμιστεί λεπτομερώς σε μια ποικιλία εργασιών με ένα ενιαίο πλαίσιο. Η αρχιτεκτονική του T5 του επιτρέπει να εκτελεί εργασίες όπως η μετάφραση, η περίληψη και ακόμη και η δημιουργία ερωτήσεων από κείμενο. Αυτή η ευελιξία καθιστά το T5 ένα ευέλικτο μοντέλο που μπορεί να προσαρμοστεί σε πολυάριθμες εφαρμογές NLP, αξιοποιώντας την ικανότητά του να χειρίζεται ποικίλους μετασχηματισμούς κειμένου (Yu et al., 2019).

2.4.2 Εφαρμογές των LLM σε εργασίες NLP

Περίληψη κειμένου

Τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ) έχουν προωθήσει σημαντικά τις δυνατότητες περίληψης κειμένου, ένα βασικό έργο στην Επεξεργασία Φυσικής Γλώσσας (ΕΠΦΓ). Η περίληψη κειμένου περιλαμβάνει τη συμπύκνωση εκτενών εγγράφων σε συντομότερες εκδόσεις που διατηρούν τις βασικές πληροφορίες και τις κύριες ιδέες. Η εφαρμογή αυτή είναι ιδιαίτερα πολύτιμη σε διάφορα επαγγελματικά πλαίσια, όπως το HR, όπου η περίληψη μακροσκελών περιγραφών θέσεων εργασίας ή προφίλ υποψηφίων μπορεί να εξοικονομήσει χρόνο και να βελτιώσει την αποτελεσματικότητα της λήψης αποφάσεων (Yu et al., 2019).

Οι LLM όπως οι GPT-3 και T5 έχουν επιδείξει αξιοσημείωτη ικανότητα στη δημιουργία συνεκτικών και συνοπτικών περιλήψεων. Η ικανότητά τους να κατανοούν και να αποστάζουν πολύπλοκες πληροφορίες σε σύντομες, ευανάγνωστες μορφές απορρέει από την εκτεταμένη εκπαίδευσή τους σε ποικίλα

κειμενικά δεδομένα. Αξιοποιώντας αυτά τα μοντέλα, οι οργανισμοί μπορούν να αυτοματοποιήσουν τη διαδικασία περίληψης εγγράφων, η οποία διαφορετικά απαιτεί σημαντική χειρωνακτική προσπάθεια (Jagwani et al., 2023).

Στο πλαίσιο της ανθρώπινου δυναμικού, η σύνοψη των περιγραφών θέσεων εργασίας με τη χρήση LLM μπορεί να βελτιώσει τη διαδικασία πρόσληψης. Οι περιγραφές θέσεων εργασίας συχνά περιέχουν λεπτομερείς πληροφορίες σχετικά με τις αρμοδιότητες, τις απαιτούμενες δεξιότητες και την εταιρική κουλτούρα. Ωστόσο, οι υποψήφιοι συχνά αναζητούν γρήγορες επισκοπήσεις για να διαπιστώσουν αν μια θέση ευθυγραμμίζεται με τα προσόντα και τα ενδιαφέροντά τους. Τα LLM μπορούν να αναλύσουν αυτές τις μακροσκελείς περιγραφές και να δημιουργήσουν συνοπτικές περιλήψεις που τονίζουν τα πιο κρίσιμα σημεία. Αυτό όχι μόνο βοηθά τους υποψηφίους να λάβουν τεκμηριωμένες αποφάσεις, αλλά και ενισχύει την προβολή των βασικών λεπτομερειών της θέσης εργασίας, προσελκύοντας ενδεχομένως περισσότερους κατάλληλους υποψηφίους (He et al., 2023).

Ομοίως, τα LLM μπορούν να χρησιμοποιηθούν για τη σύνοψη των προφίλ των υποψηφίων. Οι επαγγελματίες του ανθρώπινου δυναμικού πρέπει συχνά να εξετάζουν πολυάριθμα βιογραφικά σημειώματα και προφίλ στο LinkedIn για να εντοπίσουν πιθανές προσλήψεις. Η σύνοψη αυτών των προφίλ επιτρέπει στους υπεύθυνους προσλήψεων να αντιληφθούν γρήγορα τα προσόντα και την εμπειρία ενός υποψηφίου. Για παράδειγμα, ένας LLM μπορεί να αποστάξει ένα ολοκληρωμένο βιογραφικό σημείωμα σε μια σύντομη περίληψη που δίνει έμφαση στις βασικές δεξιότητες, τα επαγγελματικά επιτεύγματα και το σχετικό επαγγελματικό ιστορικό του υποψηφίου. Αυτή η δυνατότητα είναι ιδιαίτερα χρήσιμη όταν πρόκειται για μεγάλο όγκο αιτήσεων, διασφαλίζοντας ότι δεν παραλείπονται κρίσιμες πληροφορίες και επιτρέποντας την αποτελεσματικότερη διαλογή των υποψηφίων (Wan et al., 2023).

Μηχανική μετάφραση

Η μηχανική μετάφραση είναι ένας άλλος τομέας στον οποίο τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ) έχουν κάνει σημαντικά βήματα προόδου. Η μηχανική μετάφραση περιλαμβάνει τη μετατροπή κειμένου από μια γλώσσα σε μια άλλη, μια εργασία που απαιτεί βαθιά κατανόηση τόσο της γλώσσας προέλευσης όσο και της γλώσσας-στόχου (Nair et al., 2023). Τα LLM όπως το T5 της Google και το GPT-3 της OpenAI έχουν εκπαιδευτεί σε τεράστια πολύγλωσσα σύνολα δεδομένων, επιτρέποντάς τους να εκτελούν μεταφράσεις υψηλής ποιότητας σε πολλά ζεύγη γλωσσών (Yu et al., 2019).

Στο πλαίσιο της ανθρώπινου δυναμικού, η αυτόματη μετάφραση είναι ιδιαίτερα επωφελής για τις πολυεθνικές εταιρείες που πρέπει να επικοινωνούν τις προκηρύξεις θέσεων εργασίας και τις πληροφορίες για τους υποψηφίους σε

διάφορες γλώσσες. Η ακριβής και αποτελεσματική μετάφραση των αγγελιών εργασίας διασφαλίζει ότι οι απαιτήσεις της θέσης εργασίας γίνονται σαφώς κατανοητές από ένα ευρύ φάσμα υποψηφίων, ανεξάρτητα από τη μητρική τους γλώσσα. Αυτή η περιεκτικότητα μπορεί να επεκτείνει σημαντικά την εμβέλεια των αγγελιών εργασίας, προσελκύνοντας μια ευρύτερη και πιο ποικιλόμορφη δεξαμενή ταλέντων (Yang et al., 2023).

Για παράδειγμα, μια πολυεθνική εταιρεία μπορεί να δημοσιεύσει μια θέση εργασίας που πρέπει να διαφημιστεί σε πολλές χώρες. Χρησιμοποιώντας LLMs, η εταιρεία μπορεί να μεταφράσει την ανάρτηση εργασίας από τα αγγλικά στα γαλλικά, ισπανικά, γερμανικά και άλλες γλώσσες, εξασφαλίζοντας ότι το βασικό μήνυμα και οι απαιτήσεις παραμένουν συνεπείς σε όλες τις μεταφράσεις. Αυτή η δυνατότητα είναι απαραίτητη για τη διατήρηση της ακεραιότητας των περιγραφών θέσεων εργασίας και την αποφυγή παρεξηγήσεων που θα μπορούσαν να προκύψουν από κακώς μεταφρασμένο περιεχόμενο (Wan et al., 2023).

Επιπλέον, τα LLM μπορούν να χρησιμοποιηθούν για τη μετάφραση των προφίλ των υποψηφίων και των βιογραφικών σημειωμάτων που υποβάλλονται σε διαφορετικές γλώσσες. Σε μια παγκοσμιοποιημένη αγορά εργασίας, είναι σύνηθες οι υποψήφιοι να υποβάλλουν αίτηση για θέσεις εργασίας σε χώρες όπου η κύρια γλώσσα διαφέρει από τη δική τους. Οι LLM μπορούν να μεταφράσουν αυτά τα βιογραφικά στη γλώσσα που χρησιμοποιεί η εταιρεία πρόσληψης, διευκολύνοντας έτσι μια πιο ομαλή και χωρίς αποκλεισμούς διαδικασία πρόσληψης. Αυτό διασφαλίζει ότι οι υποψήφιοι αξιολογούνται με βάση τα προσόντα και την εμπειρία τους, αντί να μειονεκτούν λόγω γλωσσικών εμποδίων (He et al., 2023).

Δημιουργία και συμπλήρωση κειμένου

Τα Μεγάλα Γλωσσικά Μοντέλα (LLM) έχουν φέρει επανάσταση στις εργασίες δημιουργίας και συμπλήρωσης κειμένου στην Επεξεργασία Φυσικής Γλώσσας (NLP), επιτρέποντας τη δημιουργία συνεκτικού και σχετικού με το περιβάλλον κειμένου με βάση δεδομένες οδηγίες. Αυτή η ικανότητα έχει σημαντικές εφαρμογές σε διάφορους τομείς, συμπεριλαμβανομένων των ανθρώπινων πόρων (HR). Τα LLM όπως τα GPT-3 και T5 υπερέχουν στη δημιουργία και συμπλήρωση κειμένου, καθιστώντας τα ανεκτίμητα εργαλεία για τη δημιουργία αυτοματοποιημένων απαντήσεων σε chatbots και την ενίσχυση των διαδικασιών επικοινωνίας εντός των οργανισμών (Zhuo et al., 2023).

Η παραγωγή και συμπλήρωση κειμένου περιλαμβάνει την παραγωγή κειμένου που μοιάζει με ανθρώπινο κείμενο και ευθυγραμμίζεται με το πλαίσιο και το περιεχόμενο μιας δεδομένης προτροπής. Τα LLM το επιτυγχάνουν αυτό αξιοποιώντας τεράστιες ποσότητες δεδομένων εκπαίδευσης και εξελιγμένες

αρχιτεκτονικές νευρωνικών δικτύων. Για παράδειγμα, το GPT-3, με τις 175 δισεκατομμύρια παραμέτρους του, μπορεί να παράγει κείμενο που δεν είναι μόνο γραμματικά ορθό αλλά και κατάλληλο από άποψη πλαισίου και υφολογικά συνεπές με την προτροπή εισόδου. Αυτό καθιστά δυνατή τη χρήση των LLM για τη δημιουργία αυτοματοποιημένων απαντήσεων σε ερωτήματα υποψηφίων σε chatbots ανθρώπινου δυναμικού (Wan et al., 2023).

Κατά τη διαδικασία πρόσληψης, η έγκαιρη και ακριβής επικοινωνία με τους υποψηφίους είναι ζωτικής σημασίας. Τα τμήματα ανθρώπινου δυναμικού συχνά δέχονται πολυάριθμα ερωτήματα από τους υποψηφίους σχετικά με τη διαδικασία υποβολής αιτήσεων, τις λεπτομέρειες της θέσης εργασίας, τις πολιτικές της εταιρείας και πολλά άλλα (Tang et al., 2023). Η χειροκίνητη απάντηση σε κάθε ερώτημα μπορεί να είναι χρονοβόρα και επιρρεπής σε ασυνέπειες. Με την ενσωμάτωση των LLMs στα chatbots HR, οι οργανισμοί μπορούν να αυτοματοποιήσουν αυτή τη διαδικασία, παρέχοντας άμεσες και ακριβείς απαντήσεις στα ερωτήματα των υποψηφίων. Για παράδειγμα, εάν ένας υποψήφιος ρωτήσει σχετικά με την κατάσταση της αίτησής του, το chatbot μπορεί να δημιουργήσει μια απάντηση με βάση την τρέχουσα κατάσταση στο σύστημα παρακολούθησης υποψηφίων, διασφαλίζοντας ότι οι υποψήφιοι λαμβάνουν άμεσα ενημερωμένες πληροφορίες (Bhat et al., 2023).

Επιπλέον, τα LLM μπορούν να ρυθμιστούν ώστε να κατανοούν τη συγκεκριμένη γλώσσα και ορολογία που χρησιμοποιείται σε διαφορετικές βιομηχανίες και οργανισμούς, ενισχύοντας την ικανότητά τους να παράγουν σχετικές και συγκεκριμένες απαντήσεις. Αυτή η διαδικασία τελειοποίησης περιλαμβάνει την εκπαίδευση του LLM σε ένα εξειδικευμένο σύνολο δεδομένων που περιλαμβάνει παραδείγματα κοινών ερωτημάτων και κατάλληλων απαντήσεων σε έναν συγκεκριμένο τομέα. Ως αποτέλεσμα, το chatbot μπορεί να χειριστεί ένα ευρύ φάσμα ερωτήσεων, από γενικές ερωτήσεις σχετικά με την εταιρική κουλτούρα έως λεπτομερείς ερωτήσεις σχετικά με τις ευθύνες και τις απαιτήσεις της θέσης εργασίας (Yang et al., 2023).

Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος είναι μια άλλη κρίσιμη εφαρμογή των LLM στη NLP, ιδιαίτερα χρήσιμη για την ανάλυση των αξιολογήσεων και των σχολίων των εργαζομένων. Η διαδικασία αυτή περιλαμβάνει τον προσδιορισμό του συναισθήματος που εκφράζεται σε ένα κομμάτι κειμένου, όπως αν είναι θετικό, αρνητικό ή ουδέτερο. Με την αξιοποίηση των LLMs για την ανάλυση συναισθήματος, οι οργανισμοί μπορούν να αποκτήσουν πολύτιμες πληροφορίες σχετικά με την ικανοποίηση των εργαζομένων, το ηθικό στο χώρο εργασίας και τη συνολική οργανωτική υγεία (Zhuo et al., 2023).

Οι LLM όπως οι BERT και GPT-3 έχουν επιδείξει εντυπωσιακές ικανότητες στην ανάλυση συναισθήματος λόγω της βαθιάς κατανόησης του πλαισίου και των γλωσσικών αποχρώσεων. Αυτά τα μοντέλα μπορούν να ανιχνεύσουν με ακρίβεια το συναίσθημα σε διάφορες μορφές ανατροφοδότησης των εργαζομένων, συμπεριλαμβανομένων των απαντήσεων σε έρευνες, των αξιολογήσεων απόδοσης και των σχολίων στα μέσα κοινωνικής δικτύωσης. Με την ανάλυση αυτών των δεδομένων, τα τμήματα ανθρώπινου δυναμικού μπορούν να εντοπίζουν περιοχές που προκαλούν ανησυχία, να παρακολουθούν τις αλλαγές στο συναίσθημα των εργαζομένων με την πάροδο του χρόνου και να λαμβάνουν τεκμηριωμένες αποφάσεις για τη βελτίωση του εργασιακού περιβάλλοντος (Gao et al., 2023).

Για παράδειγμα, ένας οργανισμός μπορεί να χρησιμοποιήσει τα LLM για να αναλύσει τις απαντήσεις από μια έρευνα ικανοποίησης των εργαζομένων. Κατηγοριοποιώντας το συναίσθημα κάθε απάντησης, η ομάδα ανθρώπινου δυναμικού μπορεί να εντοπίσει κοινά θέματα και να εντοπίσει συγκεκριμένα ζητήματα που πρέπει να αντιμετωπιστούν (Rath et al., 2023). Εάν ένας σημαντικός αριθμός απαντήσεων υποδεικνύει δυσαρέσκεια για την επικοινωνία της διοίκησης, ο οργανισμός μπορεί να αναλάβει στοχευμένες δράσεις για τη βελτίωση της διαφάνειας και της δέσμευσης. Ομοίως, τα θετικά σχόλια για ορισμένες πτυχές του χώρου εργασίας, όπως οι ευκαιρίες για επαγγελματική ανάπτυξη, μπορούν να επισημανθούν και να ενισχυθούν (Bhat et al., 2023).

Η ανάλυση συναισθήματος παίζει επίσης καθοριστικό ρόλο στην παρακολούθηση των αξιολογήσεων των εργαζομένων σε πλατφόρμες όπως η Glassdoor και η Indeed. Αυτές οι αξιολογήσεις παρέχουν εξωτερικές προοπτικές σχετικά με την κουλτούρα, τη διοίκηση και το εργασιακό περιβάλλον της εταιρείας. Με τη χρήση των LLM για την ανάλυση του συναισθήματος αυτών των κριτικών, οι οργανισμοί μπορούν να αποκτήσουν πληροφορίες σχετικά με το πώς γίνονται αντιληπτοί από τους σημερινούς και τους πρώην εργαζομένους, να εντοπίσουν τα δυνατά και αδύνατα σημεία και να εφαρμόσουν στρατηγικές για την ενίσχυση της επωνυμίας του εργοδότη τους (Zhuo et al., 2023).

2.5 NER με LLMs

2.5.1 Πλεονεκτήματα της χρήσης LLMs για NER

Βελτιωμένη ακρίβεια και επίγνωση πλαισίου

Τα μεγάλα γλωσσικά μοντέλα (LLM) έχουν φέρει επανάσταση στην αναγνώριση ονομαστικών οντοτήτων (NER) βελτιώνοντας σημαντικά την ακρίβεια και την επίγνωση του πλαισίου. Τα παραδοσιακά συστήματα NER συχνά δυσκολεύονταν

να κατανοήσουν την πολυπλοκότητα της ανθρώπινης γλώσσας, ιδίως όταν είχαν να κάνουν με οντότητες πολλών λέξεων και ποικίλα συμφραζόμενα. Τα LLM, ωστόσο, υπερέρχουν σε αυτούς τους τομείς λόγω των προηγμένων αρχιτεκτονικών τους και της εκτεταμένης εκπαίδευσής τους σε ποικίλα σύνολα δεδομένων (Yang et al., 2023).

Η ικανότητα των LLM να αναγνωρίζουν με ακρίβεια πολύπλοκες οντότητες με πολλές λέξεις είναι ένα από τα πιο αξιοσημείωτα πλεονεκτήματά τους. Για παράδειγμα, το GPT-3, με τον τεράστιο αριθμό παραμέτρων και την εξελιγμένη αρχιτεκτονική του που βασίζεται σε μετασχηματιστές, μπορεί να αναγνωρίσει και να κατηγοριοποιήσει οντότητες όπως "Διευθύνων Σύμβουλος" ή "Ηνωμένα Έθνη" με αξιοσημείωτη ακρίβεια. Αυτό το επίπεδο ακρίβειας επιτυγχάνεται με την εκπαίδευση του μοντέλου σε ένα ποικιλόμορφο και εκτεταμένο σώμα κειμένων, το οποίο περιλαμβάνει ένα ευρύ φάσμα γλωσσικών προτύπων και συμφραζομένων. Ως αποτέλεσμα, τα LLM μπορούν να χειριστούν παραλλαγές στον τρόπο παρουσίασης των οντοτήτων, καθιστώντας τα ιδιαίτερα αποτελεσματικά σε εφαρμογές του πραγματικού κόσμου (Veen et al., 2023).

Επίγνωση του πλαισίου

Ένα άλλο σημαντικό πλεονέκτημα των LLMs στη NER είναι η ανώτερη επίγνωση του πλαισίου. Σε αντίθεση με τα παραδοσιακά μοντέλα που μπορεί να εξετάζουν μόνο ένα περιορισμένο παράθυρο του περιβάλλοντος κειμένου, τα LLM όπως το BERT (Bidirectional Encoder Representations from Transformers) αναλύουν το πλήρες πλαίσιο στο οποίο εμφανίζεται μια οντότητα. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο BERT να κατανοεί την αποχρωματισμένη σημασία των λέξεων με βάση το περιβάλλον κείμενο, διακρίνοντας μεταξύ οντοτήτων με το ίδιο όνομα αλλά διαφορετική σημασία. Για παράδειγμα, η BERT μπορεί να διακρίνει μεταξύ της "Apple" ως τεχνολογικής εταιρείας και του "μήλου" ως φρούτου, λαμβάνοντας υπόψη το ευρύτερο πλαίσιο της πρότασης. Αυτή η ικανότητα είναι ζωτικής σημασίας για την ακριβή αναγνώριση οντοτήτων σε πολύπλοκα και ποικίλα γλωσσικά περιβάλλοντα (Bishop et al., 2022).

Η βελτιωμένη ακρίβεια και η επίγνωση του πλαισίου που παρέχουν οι LLM έχουν πρακτικές συνέπειες για πολλές εφαρμογές NER. Στον τομέα των ανθρώπινων πόρων (HR), για παράδειγμα, τα μοντέλα αυτά μπορούν να εξάγουν και να κατηγοριοποιούν με ακρίβεια πληροφορίες από βιογραφικά σημειώματα, περιγραφές θέσεων εργασίας και άλλα σχετικά έγγραφα. Αυτό οδηγεί σε ακριβέστερη αντιστοίχιση των υποψηφίων με τις απαιτήσεις της θέσης εργασίας, ενισχύοντας την αποδοτικότητα και την αποτελεσματικότητα της διαδικασίας πρόσληψης. Ομοίως, στα χρηματοοικονομικά, τα μοντέλα LLM μπορούν να αναγνωρίζουν και να κατηγοριοποιούν με ακρίβεια οντότητες σε χρηματοοικονομικά έγγραφα, όπως ονόματα εταιρειών, λεπτομέρειες

συναλλαγών και κανονιστικούς όρους, διευκολύνοντας την καλύτερη ανάλυση δεδομένων και τη λήψη αποφάσεων (Xie et al., 2023).

Επεκτασιμότητα

Εκτός από τη βελτιωμένη ακρίβεια και την επίγνωση του πλαισίου, οι LLM προσφέρουν σημαντικά πλεονεκτήματα κλιμάκωσης για εργασίες NER. Η ικανότητά τους να χειρίζονται μεγάλα σύνολα δεδομένων και ποικίλους τύπους κειμένου τα καθιστά ιδανικά για την αποτελεσματική επεξεργασία τεράστιων όγκων πληροφοριών. Αυτή η επεκτασιμότητα είναι ιδιαίτερα σημαντική σε κλάδους που παράγουν ή διαχειρίζονται εκτεταμένα δεδομένα κειμένου, όπως οι τομείς της ανθρώπινου δυναμικού, της οικονομίας, της υγειονομικής περίθαλψης και των νομικών υπηρεσιών (Ushio & Camacho-Collados, 2022).

Οι LLM που βασίζονται στο νέφος αποτελούν παράδειγμα αυτής της επεκτασιμότητας, επιτρέποντας στους οργανισμούς να επεξεργάζονται παράλληλα εκατομμύρια έγγραφα χωρίς συμβιβασμούς στην απόδοση. Για παράδειγμα, οι υλοποιήσεις της GPT-3 που βασίζονται στο νέφος μπορούν να αναλύσουν και να εξάγουν οντότητες από τεράστιο όγκο βιογραφικών σημειωμάτων, αιτήσεων εργασίας και αρχείων εργαζομένων. Αυτή η δυνατότητα επιτρέπει στα τμήματα ανθρώπινου δυναμικού να εξορθολογίζουν τις ροές εργασίας τους, μειώνοντας τον χρόνο και τους πόρους που απαιτούνται για τη χειροκίνητη καταχώρηση και αναθεώρηση δεδομένων. Αξιοποιώντας την υποδομή cloud, οι οργανισμοί μπορούν επίσης να κλιμακώνουν δυναμικά την ικανότητα επεξεργασίας τους για να αντιμετωπίζουν φορτία αιχμής, όπως κατά τη διάρκεια κινήσεων πρόσληψης ή κανονιστικών ελέγχων (Bian et al., 2023).

Επιπλέον, η επεκτασιμότητα των LLM επεκτείνεται στην προσαρμοστικότητα τους σε διαφορετικούς τομείς και γλώσσες. Οι LLM όπως η BERT και η GPT-3 μπορούν να ρυθμιστούν λεπτομερώς σε σύνολα δεδομένων συγκεκριμένων τομέων για να βελτιώσουν την απόδοσή τους σε εξειδικευμένους τομείς. Για παράδειγμα, ένα νομικό γραφείο μπορεί να προσαρμόσει ένα LLM σε νομικά κείμενα για να βελτιώσει την ακρίβειά του στον εντοπισμό νομικών οντοτήτων, αναφορών σε υποθέσεις και θεσμοθετημένων όρων. Παρομοίως, μια πολυεθνική εταιρεία θα μπορούσε να τελειοποιήσει ένα LLM σε πολύγλωσσα σύνολα δεδομένων για να υποστηρίξει την αναγνώριση οντοτήτων σε πολλές γλώσσες, διευκολύνοντας τις παγκόσμιες λειτουργίες και την επικοινωνία (Xu et al., 2021).

Η ενσωμάτωση των LLM με άλλες προηγμένες τεχνολογίες ενισχύει περαιτέρω την επεκτασιμότητα και τη χρησιμότητά τους. Για παράδειγμα, ο συνδυασμός των LLM με την τεχνολογία οπτικής αναγνώρισης χαρακτήρων (OCR) επιτρέπει στους οργανισμούς να εξάγουν οντότητες από σαρωμένα έγγραφα και εικόνες, διευρύνοντας το φάσμα των πηγών που μπορούν να υποβληθούν σε αυτόματη επεξεργασία. Αυτή η ενσωμάτωση είναι ιδιαίτερα πολύτιμη σε τομείς που

ασχολούνται με μεγάλο όγκο φυσικών εγγράφων, όπως η υγειονομική περίθαλψη και τα οικονομικά, όπου η ψηφιοποίηση και η ανάλυση των έντυπων αρχείων μπορεί να οδηγήσει σε σημαντική αύξηση της αποδοτικότητας (Vijay & Priyanshu, 2022).

Παρά τα πολλά πλεονεκτήματα, είναι σημαντικό να αντιμετωπιστούν οι προκλήσεις που σχετίζονται με την επεκτασιμότητα των LLM. Οι υπολογιστικοί πόροι που απαιτούνται για την εκπαίδευση και την ανάπτυξη αυτών των μοντέλων μπορεί να είναι σημαντικοί, γεγονός που απαιτεί επενδύσεις σε υλικό υψηλής απόδοσης και υποδομές νέφους. Επιπλέον, η διαχείριση του απορρήτου και της ασφάλειας των δεδομένων είναι ζωτικής σημασίας κατά την επεξεργασία ευαίσθητων πληροφοριών σε κλίμακα. Οι οργανισμοί πρέπει να εφαρμόζουν ισχυρές πρακτικές διακυβέρνησης δεδομένων για να διασφαλίζουν τη συμμόρφωση με τους κανονισμούς και να προστατεύονται από παραβιάσεις δεδομένων (Amalvy et al., 2023).

2.5.2 Μελέτες περιπτώσεων και εφαρμογές

Μελέτη περίπτωσης 1: Ακαδημαϊκή έρευνα σχετικά με τη NEP με τη χρήση LLMs

Η ακαδημαϊκή έρευνα έχει συμβάλει καθοριστικά στην απόδειξη της αποτελεσματικότητας των μεγάλων γλωσσικών μοντέλων (LLM) για την αναγνώριση ονομαστικών οντοτήτων (NER). Μια αξιοσημείωτη μελέτη επικεντρώθηκε στη λεπτομερή ρύθμιση των μοντέλων BERT για εργασίες NER, αναδεικνύοντας την ικανότητα του μοντέλου να ξεπερνά σημαντικά τις παραδοσιακές προσεγγίσεις NER. Οι ερευνητές διεξήγαγαν εκτεταμένα πειράματα με τη χρήση του BERT, αξιοποιώντας την αρχιτεκτονική του αμφίδρομου μετασχηματιστή για τη σύλληψη του πλαισίου και από τις δύο κατευθύνσεις σε μια ακολουθία κειμένου. Αυτή η αμφίδρομη προσέγγιση επιτρέπει στο BERT να κατανοεί το πλήρες πλαίσιο κάθε λέξης, οδηγώντας σε ακριβέστερη αναγνώριση οντοτήτων (Sajid et al., 2022).

Στη μελέτη, οι ερευνητές τελειοποίησαν το BERT σε διάφορα σύνολα δεδομένων NER, συμπεριλαμβανομένων των CoNLL-2003 και OntoNotes, τα οποία περιέχουν σχολιασμένο κείμενο για οντότητες όπως ονόματα προσώπων, οργανισμούς και τοποθεσίες. Η διαδικασία τελειοποίησης περιελάμβανε την εκπαίδευση του προ-εκπαιδευμένου μοντέλου BERT σε αυτά τα συγκεκριμένα σύνολα δεδομένων NER, επιτρέποντάς του να προσαρμοστεί στις αποχρώσεις και τα μοτίβα που είναι μοναδικά στις εργασίες αναγνώρισης οντοτήτων. Τα αποτελέσματα ήταν εντυπωσιακά, με το μοντέλο BERT που έχει ρυθμιστεί με ακρίβεια να επιτυγχάνει κορυφαίες μετρήσεις επιδόσεων, όπως υψηλότερες βαθμολογίες ακρίβειας, ανάκλησης και F1 σε σύγκριση με προηγούμενα μοντέλα (Du et al., 2023).

Η μελέτη ανέδειξε επίσης την ευελιξία του BERT στο χειρισμό διαφορετικών τύπων οντοτήτων και διαφορετικών πηγών κειμένου. Με τη λεπτομερή ρύθμιση του BERT σε σύνολα δεδομένων συγκεκριμένων τομέων, το μοντέλο μπορούσε να αναγνωρίζει αποτελεσματικά εξειδικευμένες οντότητες, όπως ιατρικούς όρους σε κλινικά κείμενα ή νομικές οντότητες σε νομικά έγγραφα. Αυτή η προσαρμοστικότητα καθιστά το BERT ένα ισχυρό εργαλείο για διάφορες εφαρμογές NER σε διάφορους κλάδους. Η έρευνα κατέληξε στο συμπέρασμα ότι τα λεπτομερώς ρυθμισμένα μοντέλα BERT θα μπορούσαν να χρησιμεύσουν ως ένα ισχυρό θεμέλιο για την ανάπτυξη προηγμένων συστημάτων NER, προσφέροντας ανώτερη ακρίβεια και επίγνωση του πλαισίου (Ashok & Lipton, 2023).

Μελέτη περίπτωσης 2: Έργα ανοικτού κώδικα που χρησιμοποιούν LLMs για NER

Η κοινότητα ανοικτού κώδικα έχει αγκαλιάσει ενεργά τα μοντέλα LLM για NER, με πολυάριθμα έργα σε πλατφόρμες όπως το GitHub που παρουσιάζουν καινοτόμες χρήσεις μοντέλων όπως το BERT και το GPT-3 για την εξαγωγή οντοτήτων. Αυτά τα έργα παρέχουν πολύτιμους πόρους και εργαλεία για προγραμματιστές και ερευνητές ώστε να υλοποιήσουν και να πειραματιστούν με NER χρησιμοποιώντας LLMs τελευταίας τεχνολογίας (Jia et al., 2020).

Ένα χαρακτηριστικό παράδειγμα είναι η βιβλιοθήκη spaCy, η οποία ενσωματώνει προ-εκπαιδευμένα μοντέλα BERT για εργασίες NER. Ο χαρακτήρας ανοικτού κώδικα της spaCy επιτρέπει στους προγραμματιστές να έχουν εύκολη πρόσβαση και να χρησιμοποιούν τις προηγμένες δυνατότητες NER στις εφαρμογές τους. Η βιβλιοθήκη παρέχει ένα ευέλικτο API για τη λεπτομερή ρύθμιση των μοντέλων BERT σε προσαρμοσμένα σύνολα δεδομένων, επιτρέποντας στους χρήστες να προσαρμόσουν το σύστημα NER στις συγκεκριμένες ανάγκες τους. Η ενσωμάτωση του BERT στο SpaCy έχει υιοθετηθεί ευρέως σε διάφορα έργα, από την ακαδημαϊκή έρευνα έως τις βιομηχανικές εφαρμογές, αποδεικνύοντας την αποτελεσματικότητα και την ευκολία χρήσης του (Li et al., 2023).

Ένα άλλο αξιοσημείωτο έργο ανοικτού κώδικα είναι η βιβλιοθήκη Hugging Face Transformers, η οποία προσφέρει υλοποιήσεις πολλών LLMs, συμπεριλαμβανομένων των BERT, GPT-3 και T5. Αυτή η βιβλιοθήκη παρέχει προ-εκπαιδευμένα μοντέλα και εργαλεία για λεπτομερή ρύθμιση σε σύνολα δεδομένων NER, διευκολύνοντας την ανάπτυξη προσαρμοσμένων συστημάτων NER. Η κοινότητα Hugging Face έχει συνεισφέρει εκτενή τεκμηρίωση, σεμινάρια και παραδείγματα, καθιστώντας την προσιτή σε προγραμματιστές όλων των επιπέδων δεξιοτήτων. Τα έργα που χρησιμοποιούν μετασχηματιστές Hugging Face για NER έχουν παρουσιάσει σημαντικές βελτιώσεις στην απόδοση αναγνώρισης οντοτήτων, επωφελούμενα από τις προηγμένες δυνατότητες των LLMs (Rasmy et al., 2020).

Αυτές οι πρωτοβουλίες ανοικτού κώδικα αναδεικνύουν τη συλλογική προσπάθεια για την προώθηση της τεχνολογίας NER με τη χρήση LLM. Με την παροχή προσβάσιμων εργαλείων και πόρων, η κοινότητα ανοικτού κώδικα επιτρέπει τον εκτεταμένο πειραματισμό και την καινοτομία, οδηγώντας σε περαιτέρω βελτιώσεις στις εφαρμογές NER (Jia et al., 2020).

Πραγματικές εφαρμογές

Στη βιομηχανία, οι LLM έχουν εφαρμοστεί με επιτυχία σε διάφορες πραγματικές εφαρμογές, αποδεικνύοντας την πρακτική αξία και την αποτελεσματικότητά τους. Για παράδειγμα, μεγάλες πύλες εργασίας έχουν ενσωματώσει τις LLM στις πλατφόρμες τους για να βελτιώσουν τους αλγορίθμους αντιστοίχισης θέσεων εργασίας, βελτιώνοντας σημαντικά τη διαδικασία πρόσληψης (Ushio & Camacho-Collados, 2022).

Το LinkedIn είναι ένα χαρακτηριστικό παράδειγμα μιας πύλης εργασίας που αξιοποιεί τα LLMs για NER. Το LinkedIn χρησιμοποιεί προηγμένα μοντέλα όπως το BERT για την ανάλυση προφίλ χρηστών, δημοσιεύσεων θέσεων εργασίας και άλλων σχετικών εγγράφων. Εξάγοντας και κατηγοριοποιώντας οντότητες όπως τίτλους θέσεων εργασίας, δεξιότητες και ονόματα εταιρειών, το σύστημα NER του LinkedIn μπορεί να αντιστοιχίσει με ακρίβεια τους υποψηφίους με τις ευκαιρίες απασχόλησης. Αυτό έχει οδηγήσει σε πιο σχετικές συστάσεις θέσεων εργασίας για τους χρήστες και πιο αποτελεσματικές αναζητήσεις υποψηφίων για τους υπεύθυνους προσλήψεων, ενισχύοντας τη συνολική ικανοποίηση και δέσμευση των χρηστών (Zhou et al., 2023).

Μια άλλη επιτυχημένη εφαρμογή μπορεί να παρατηρηθεί στις χρηματοπιστωτικές υπηρεσίες, όπου οι εταιρείες χρησιμοποιούν LLM για την επεξεργασία και την ανάλυση οικονομικών εγγράφων. Για παράδειγμα, το Bloomberg έχει ενσωματώσει LLMs στον αγωγό επεξεργασίας δεδομένων του για την εξαγωγή οντοτήτων από οικονομικές εκθέσεις, άρθρα ειδήσεων και αναλύσεις αγοράς. Αυτοματοποιώντας την εξαγωγή κρίσιμων πληροφοριών, όπως ονόματα εταιρειών, χρηματοοικονομικές μετρήσεις και κανονιστικούς όρους, το Bloomberg μπορεί να παρέχει έγκαιρες και ακριβείς πληροφορίες στους χρήστες του, βελτιώνοντας τις διαδικασίες λήψης αποφάσεων (Ashok & Lipton, 2023).

Η υγειονομική περίθαλψη είναι ένας άλλος τομέας που επωφελείται από τα LLMs για το NER. Εταιρείες όπως η IBM Watson Health χρησιμοποιούν LLMs για να εξάγουν οντότητες από ιατρικά αρχεία, ερευνητικά έγγραφα και εκθέσεις κλινικών δοκιμών. Με τον ακριβή εντοπισμό ιατρικών όρων, πληροφοριών για τον ασθενή και λεπτομερειών θεραπείας, τα συστήματα αυτά βελτιώνουν τη διαχείριση και την ανάλυση δεδομένων, οδηγώντας σε καλύτερη περίθαλψη των ασθενών και αποτελεσματικότερη κλινική έρευνα (Zhou et al., 2023).

Αυτές οι πραγματικές υλοποιήσεις καταδεικνύουν τον μετασχηματιστικό αντίκτυπο των LLMs στην NER σε διάφορες βιομηχανίες. Παρέχοντας υψηλότερη ακρίβεια, επίγνωση του πλαισίου και επεκτασιμότητα, οι LLMs επιτρέπουν στους οργανισμούς να αυτοματοποιήσουν σύνθετες εργασίες επεξεργασίας κειμένου, να βελτιώσουν τις ροές εργασίας και να αποκτήσουν βαθύτερες γνώσεις από τα δεδομένα τους. Καθώς η τεχνολογία LLM συνεχίζει να εξελίσσεται, οι εφαρμογές της στην NER θα επεκταθούν πιθανότατα, οδηγώντας σε περαιτέρω καινοτομίες και βελτιώσεις στην επεξεργασία και ανάλυση δεδομένων σε διάφορους τομείς (Ushio & Camacho-Collados, 2022).

2.5.3 Προκλήσεις και μελλοντικές κατευθύνσεις

Υπολογιστικές απαιτήσεις

Η ανάπτυξη μεγάλων γλωσσικών μοντέλων (LLM) για την αναγνώριση ονομαστικών οντοτήτων (NER) συνεπάγεται σημαντικές υπολογιστικές απαιτήσεις, θέτοντας σημαντικές προκλήσεις για τους ερευνητές και τους επαγγελματίες. Η εκπαίδευση και η λεπτομερής ρύθμιση των LLM, όπως το GPT-3 και το BERT, απαιτούν τεράστιους υπολογιστικούς πόρους, συμπεριλαμβανομένου υλικού υψηλής απόδοσης και εκτεταμένης κατανάλωσης ενέργειας. Αυτά τα μοντέλα, με δισεκατομμύρια παραμέτρους, απαιτούν ισχυρές GPU ή TPU για να χειριστούν τους πολύπλοκους υπολογισμούς που εμπλέκονται στις λειτουργίες τους. Αυτός ο απαιτητικός σε πόρους χαρακτήρας των LLM μπορεί να δημιουργήσει εμπόδια στην ευρεία υιοθέτησή τους, ιδίως για οργανισμούς με περιορισμένη πρόσβαση σε προηγμένη υπολογιστική υποδομή (Prasad et al., 2023).

Η εκπαίδευση ενός LLM όπως το GPT-3 είναι ένα μνημειώδες έργο που περιλαμβάνει την επεξεργασία τεράστιων συνόλων δεδομένων μέσω πολλών επιπέδων νευρωνικών δικτύων. Η διαδικασία αυτή όχι μόνο απαιτεί σημαντικό χρόνο αλλά και συνεπάγεται υψηλό ενεργειακό κόστος, συμβάλλοντας στο περιβαλλοντικό αποτύπωμα της έρευνας τεχνητής νοημοσύνης. Ο τεράστιος όγκος των δεδομένων και η ανάγκη για επαναληπτικούς κύκλους βελτιστοποίησης σημαίνουν ότι οι συνεδρίες εκπαίδευσης μπορεί να διαρκέσουν εβδομάδες ή και μήνες. Για παράδειγμα, η εκπαίδευση του GPT-3 από το OpenAI φέρεται να κατανάλωσε petaflops υπολογιστικής ισχύος για παρατεταμένες περιόδους, υπογραμμίζοντας τους τεράστιους πόρους που απαιτούνται για τέτοιες προσπάθειες (Journal, 2023).

Η ανάπτυξη αυτών των μοντέλων σε πραγματικές εφαρμογές θέτει επίσης υπολογιστικές προκλήσεις. Η εκτέλεση LLM για εργασίες NER σε περιβάλλοντα παραγωγής απαιτεί ισχυρή και κλιμακούμενη υποδομή ικανή να υποστηρίζει

επεξεργασία υψηλής απόδοσης και χαμηλής καθυστέρησης. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές όπως η ανάλυση βιογραφικών σημειωμάτων σε πραγματικό χρόνο ή η ανάλυση εγγράφων μεγάλης κλίμακας, όπου η απόδοση και η ταχύτητα είναι κρίσιμες. Οι οργανισμοί πρέπει να επενδύσουν σε λύσεις που βασίζονται σε cloud ή σε ειδικό υλικό για να διασφαλίσουν ότι τα συστήματα NER μπορούν να λειτουργούν αποτελεσματικά και να διαχειρίζονται μεγάλους όγκους δεδομένων χωρίς σημαντικές καθυστερήσεις (Julian & Subeka, 2023).

Επιπλέον, οι υπολογιστικές απαιτήσεις των LLM έχουν επιπτώσεις στην προσβασιμότητα και την ισότητα στην έρευνα TN (Li et al., 2020). Τα μικρότερα ερευνητικά ιδρύματα, οι νεοσύστατες επιχειρήσεις και οι εταιρείες σε αναπτυσσόμενες περιοχές μπορεί να δυσκολεύονται να αντέξουν οικονομικά το απαραίτητο υλικό και τις υπηρεσίες cloud, περιορίζοντας τη δυνατότητά τους να αξιοποιήσουν τα LLM τελευταίας τεχνολογίας. Αυτή η ανισότητα μπορεί να διευρύνει το χάσμα μεταξύ των καλά χρηματοδοτούμενων φορέων και εκείνων με λιγότερους πόρους, καταπνίγοντας ενδεχομένως την καινοτομία και τη συνεργασία σε ολόκληρη την κοινότητα της TN (Klein et al., 2023).

Ανησυχίες σχετικά με το απόρρητο των δεδομένων

Η χρήση των LLMs για NER εγείρει επίσης σημαντικές ανησυχίες σχετικά με την προστασία της ιδιωτικής ζωής των δεδομένων, ιδίως όταν τα μοντέλα αυτά επεξεργάζονται ευαίσθητες ή προσωπικές πληροφορίες. Δεδομένου ότι τα LLM συχνά εκπαιδεύονται σε τεράστια σύνολα δεδομένων που μπορεί να περιλαμβάνουν ιδιότητα ή εμπιστευτικά δεδομένα, η διασφάλιση της συμμόρφωσης με τους κανονισμούς προστασίας δεδομένων είναι υψίστης σημασίας. Μοντέλα όπως το GPT-3 και το BERT, όταν εφαρμόζονται σε εργασίες που αφορούν προσωπικά δεδομένα, πρέπει να τηρούν αυστηρά πρότυπα προστασίας της ιδιωτικής ζωής για την προστασία των δικαιωμάτων των ατόμων και την αποτροπή μη εξουσιοδοτημένης πρόσβασης ή κατάχρησης των πληροφοριών (Agostinelli et al., 2019).

Μία από τις κύριες ανησυχίες είναι ο κίνδυνος ακούσιας διαρροής δεδομένων κατά τη διαδικασία εκπαίδευσης. Οι LLM μπορούν να απομνημονεύσουν και να αναπαράγουν ακούσια συγκεκριμένα σημεία δεδομένων από τα σύνολα εκπαίδευσης, τα οποία μπορεί να περιλαμβάνουν ευαίσθητες πληροφορίες, όπως ονόματα, διευθύνσεις ή οικονομικά στοιχεία. Το ζήτημα αυτό είναι ιδιαίτερα προβληματικό σε τομείς όπως η υγειονομική περίθαλψη και τα οικονομικά, όπου η διατήρηση της εμπιστευτικότητας των αρχείων ασθενών ή των οικονομικών συναλλαγών είναι κρίσιμη. Οι ερευνητές και οι επαγγελματίες πρέπει να εφαρμόζουν ισχυρές τεχνικές ανωνυμοποίησης και αποταυτοποίησης δεδομένων

για να ελαχιστοποιήσουν τον κίνδυνο έκθεσης προσωπικών δεδομένων κατά την εκπαίδευση και την ανάπτυξη μοντέλων (Caruccio et al., 2020).

Η διασφάλιση της συμμόρφωσης με τους κανονισμούς προστασίας δεδομένων, όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση, είναι μια άλλη κρίσιμη πτυχή της χρήσης των LLMs για NER. Ο ΓΚΠΔ επιβάλλει αυστηρές απαιτήσεις σχετικά με την επεξεργασία, την αποθήκευση και τη διαβίβαση δεδομένων προσωπικού χαρακτήρα, επιβάλλοντας στους οργανισμούς να εφαρμόζουν κατάλληλα τεχνικά και οργανωτικά μέτρα για τη διασφάλιση της ιδιωτικής ζωής. Αυτό περιλαμβάνει τη λήψη ρητής συγκατάθεσης από τα άτομα για την επεξεργασία δεδομένων, τη διασφάλιση της ελαχιστοποίησης των δεδομένων και την παροχή μηχανισμών για τα άτομα να έχουν πρόσβαση, να διορθώνουν ή να διαγράφουν τα δεδομένα τους (Wang et al., 2019).

Οι οργανισμοί που χρησιμοποιούν LLMs για NER πρέπει επίσης να επαγρυπνούν σχετικά με την ασφάλεια των δεδομένων. Η εφαρμογή ισχυρών πρωτοκόλλων κρυπτογράφησης για την αποθήκευση και τη μετάδοση δεδομένων είναι απαραίτητη για την προστασία από παραβιάσεις και μη εξουσιοδοτημένη πρόσβαση. Οι τακτικοί έλεγχοι ασφάλειας και συμμόρφωσης μπορούν να βοηθήσουν στον εντοπισμό και την αντιμετώπιση πιθανών τρωτών σημείων του συστήματος. Επιπλέον, η διαφάνεια στις πρακτικές επεξεργασίας δεδομένων, συμπεριλαμβανομένης της σαφούς επικοινωνίας με τους χρήστες σχετικά με τον τρόπο χρήσης και προστασίας των δεδομένων τους, μπορεί να οικοδομήσει εμπιστοσύνη και να διασφαλίσει την τήρηση ηθικών προτύπων (Caruccio et al., 2020).

Δεοντολογικές εκτιμήσεις

Η ενσωμάτωση των μεγάλων γλωσσικών μοντέλων (LLM) στην αναγνώριση ονομαστικών οντοτήτων (NER) στο πλαίσιο διαδικασιών ανθρώπινου δυναμικού (HR) εγείρει διάφορες ηθικές πτυχές που πρέπει να αντιμετωπιστούν σχολαστικά για να διασφαλιστούν δίκαια και αμερόληπτα αποτελέσματα. Η πρωταρχική ηθική πρόκληση έγκειται στον μετριασμό της προκατάληψης στις συστάσεις και τις αποφάσεις που παράγονται από LLM. Τα LLM, όπως το GPT-3 και το BERT, εκπαιδεύονται σε εκτεταμένα σύνολα δεδομένων που μπορεί να αντανακλούν ακούσια τις κοινωνικές προκαταλήψεις που υπάρχουν στα δεδομένα. Κατά συνέπεια, αυτά τα μοντέλα μπορούν να διαιωνίσουν και ακόμη και να επιδεινώσουν τις υπάρχουσες προκαταλήψεις στις διαδικασίες ανθρώπινου δυναμικού, οδηγώντας σε άδικη μεταχείριση των υποψηφίων και των εργαζομένων (Agostinelli et al., 2019).

Η προκατάληψη στα LLM μπορεί να εκδηλωθεί με διάφορους τρόπους. Για παράδειγμα, κατά την ανάλυση βιογραφικών σημειωμάτων και τη δημιουργία

συστάσεων υποψηφίων, ένα LLM μπορεί να ευνοεί ορισμένες δημογραφικές ομάδες έναντι άλλων με βάση τα πρότυπα που έχει μάθει από ιστορικά δεδομένα προσλήψεων. Αυτό θα μπορούσε να οδηγήσει σε πρακτικές διακρίσεων, όπως η μεροληψία λόγω φύλου ή φυλής, όπου οι υποψήφιοι από υποεκπροσωπούμενες ομάδες μειονεκτούν συστηματικά. Για να αντιμετωπιστεί αυτό, είναι ζωτικής σημασίας να εφαρμοστούν στρατηγικές για τον εντοπισμό και τον μετριασμό της προκατάληψης στο πλαίσιο των LLM. Τεχνικές όπως η εκπαίδευση με επίγνωση της δικαιοσύνης, όπου τα μοντέλα εκπαιδεύονται σε ισορροπημένα σύνολα δεδομένων και υπόκεινται σε περιορισμούς δικαιοσύνης, μπορούν να βοηθήσουν στη μείωση της προκατάληψης. Επιπλέον, η συνεχής παρακολούθηση και ο έλεγχος των αποτελεσμάτων των μοντέλων είναι απαραίτητοι για να διασφαλιστεί ότι οι συστάσεις παραμένουν δίκαιες και αμερόληπτες με την πάροδο του χρόνου (Caruccio et al., 2020).

Η διαφάνεια στη χρήση των LLMs για NER στο HR είναι μια άλλη κρίσιμη δεοντολογική θεώρηση. Οι οργανισμοί πρέπει να γνωστοποιούν με σαφήνεια στους υποψηφίους και τους εργαζομένους τον τρόπο με τον οποίο χρησιμοποιούνται αυτά τα μοντέλα στις διαδικασίες λήψης αποφάσεων. Αυτό περιλαμβάνει την εξήγηση του ρόλου των LLMs στην ανάλυση βιογραφικών σημειωμάτων, στην αντιστοίχιση θέσεων εργασίας και σε άλλες λειτουργίες ανθρώπινου δυναμικού (Ayala-Rivera & Pasquale, 2018). Με την προώθηση της διαφάνειας, οι οργανισμοί μπορούν να οικοδομήσουν εμπιστοσύνη και να διασφαλίσουν ότι τα άτομα κατανοούν τον τρόπο με τον οποίο χρησιμοποιούνται τα δεδομένα τους και τις επιπτώσεις της αυτοματοποιημένης λήψης αποφάσεων. Επιπλέον, η παροχή μηχανισμών για τα άτομα να αμφισβητούν και να ζητούν διευκρινίσεις σχετικά με τις αποφάσεις που λαμβάνονται από τα LLMs είναι απαραίτητη για τη διατήρηση της λογοδοσίας και της δικαιοσύνης (Wang et al., 2019).

Μελλοντικές ερευνητικές κατευθύνσεις

Η μελλοντική έρευνα στον τομέα των LLMs και των NER θα πρέπει να επικεντρωθεί σε διάφορους βασικούς τομείς για την ενίσχυση των δυνατοτήτων και της ηθικής εφαρμογής αυτών των μοντέλων. Ένας σημαντικός τομέας είναι η βελτίωση της ερμηνευσιμότητας των μοντέλων (Barati & Rana, 2020). Τα τρέχοντα LLM λειτουργούν ως "μαύρα κουτιά", καθιστώντας δύσκολη την κατανόηση των διαδικασιών λήψης αποφάσεων που βρίσκονται πίσω από τα αποτελέσματά τους. Η βελτίωση της ερμηνευσιμότητας περιλαμβάνει την ανάπτυξη μεθόδων που επιτρέπουν στους χρήστες να εντοπίζουν και να εξηγούν πώς λαμβάνονται συγκεκριμένες αποφάσεις από το μοντέλο. Αυτή η διαφάνεια είναι ζωτικής σημασίας για τη διασφάλιση της εμπιστοσύνης και της λογοδοσίας

στα αυτοματοποιημένα συστήματα, ιδίως σε ευαίσθητους τομείς όπως το HR (Capodiecici & Mainetti, 2020).

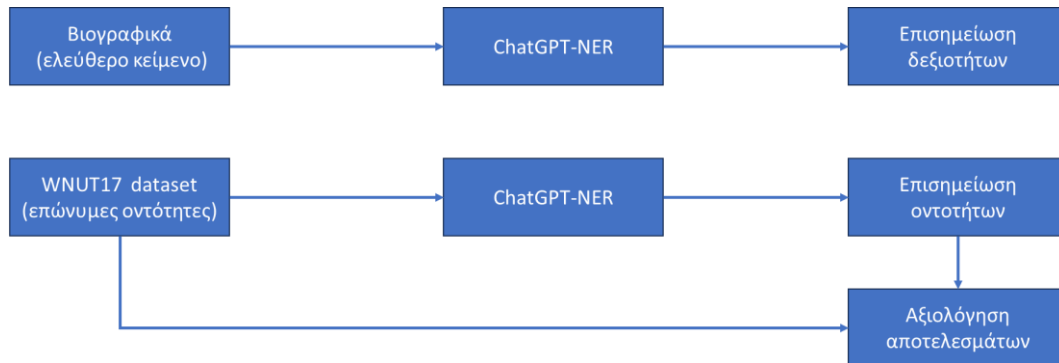
Οι ερευνητικές προσπάθειες θα πρέπει επίσης να στοχεύουν στη μείωση των προκαταλήψεων στα LLM. Παρά τις υπάρχουσες τεχνικές για τον μετριασμό της μεροληψίας, η επίτευξη πλήρους δικαιοσύνης παραμένει ένας δύσκολος στόχος. Η ανάπτυξη πιο ισχυρών αλγορίθμων που μπορούν να ανιχνεύουν και να εξαλείφουν τις λεπτές μεροληψίες στα δεδομένα εκπαίδευσης και στις εξόδους του μοντέλου είναι ζωτικής σημασίας. Αυτό περιλαμβάνει τη δημιουργία συνόλων δεδομένων που είναι αντιπροσωπευτικά για διαφορετικούς πληθυσμούς και το σχεδιασμό μοντέλων που μπορούν να προσαρμόζονται για τυχόν ανισορροπίες στα δεδομένα. Οι συνεργατικές προσπάθειες μεταξύ των ερευνητών TN, των ηθικολόγων και των εμπειρογνομόνων του τομέα είναι απαραίτητες για την εκπόνηση ολοκληρωμένων στρατηγικών για τον μετριασμό των προκαταλήψεων (Talboy & Fuller, 2023).

Μια άλλη πολλά υποσχόμενη κατεύθυνση για μελλοντική έρευνα είναι η ενσωμάτωση των LLMs με άλλες τεχνικές TN για τη βελτίωση της απόδοσης και της δυνατότητας εφαρμογής της NER. Για παράδειγμα, ο συνδυασμός LLMs με γράφους γνώσης μπορεί να βελτιώσει την ακρίβεια και την κατανόηση του πλαισίου της αναγνώρισης οντοτήτων παρέχοντας δομημένες πληροφορίες υποβάθρου. Αυτή η υβριδική προσέγγιση μπορεί να καταστήσει δυνατή την εξαγωγή οντοτήτων με περισσότερες αποχρώσεις και επίγνωση του πλαισίου, γεγονός που είναι ιδιαίτερα επωφελές σε σύνθετα σενάρια ανθρώπινου δυναμικού (Huang et al., 2023).

Η επεκτασιμότητα και η αποδοτικότητα των LLM δικαιολογούν επίσης περαιτέρω διερεύνηση. Η εκπαίδευση και η ανάπτυξη των LLM είναι διαδικασίες έντασης πόρων που μπορεί να περιορίσουν την προσβασιμότητά τους. Η έρευνα για αποδοτικότερους αλγορίθμους εκπαίδευσης και βελτιστοποιήσεις υλικού μπορεί να συμβάλει στο να γίνουν οι LLM πιο προσιτές και βιώσιμες. Αυτό περιλαμβάνει τη διερεύνηση τεχνικών όπως η απόσταξη μοντέλων, η οποία περιλαμβάνει την εκπαίδευση μικρότερων, πιο αποδοτικών μοντέλων που διατηρούν την απόδοση μεγαλύτερων αντίστοιχων μοντέλων (Mohammad et al., 2021).

3. Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε συνοψίζεται στο σχήμα 1.



Σχήμα 1. Η μεθοδολογία ανάλυσης. Αρχικά, δοκιμάζουμε την μεθοδολογία ChatGPT-NER σε δεδομένα βιογραφικών για εύρεση δεξιοτήτων. Στη συνέχεια, χρησιμοποιήσαμε το σύνολο δεδομένων WNUT17, το οποίο έχει ήδη σημειωμένες τις επώνυμες οντότητες, τις οποίες χρησιμοποιούμε για αξιολόγηση των αποτελεσμάτων της μεθόδου ChatGPT-NER.

3.1. Μοντέλο GPT-NER

Παρά το γεγονός ότι τα μοντέλα γλώσσας μεγάλης κλίμακας (LLM) έχουν επιτύχει υψηλές επιδόσεις σε μια ποικιλία εργασιών NLP, η απόδοσή τους στο NER εξακολουθεί να είναι σημαντικά χαμηλότερη από βασικές μεθόδους. Αυτό οφείλεται στο χάσμα μεταξύ των δύο εργασιών NER και LLM: η πρώτη είναι μια εργασία επισήμανσης ακολουθίας, ενώ η δεύτερη είναι ένα μοντέλο δημιουργίας κειμένου.

Στην εργασία των Wang et al., (2023) προτάθηκε το GPT-NER για την επίλυση αυτού του ζητήματος. Το GPT-NER γεφυρώνει το χάσμα μετατρέποντας την εργασία επισήμανσης ακολουθίας σε μια εργασία δημιουργίας που μπορεί εύκολα να εκτελεστεί από LLM, π.χ., η εργασία εύρεσης οντοτήτων τοποθεσίας στο κείμενο εισαγωγής «Η Αθήνα είναι μια πόλη» μετατρέπεται για να δημιουργήσει την ακολουθία κειμένου «Η @@Αθήνα## είναι μια πόλη», όπου τα ειδικά διακριτικά @@## σηματοδοτούν την οντότητα προς εξαγωγή. Ένα πλεονέκτημα αυτής της μεθόδου είναι ότι μπορεί να δουλέψει με μικρό σύνολο δεδομένων, σε αντίθεση με άλλα μοντέλα που βασίζονται σε νευρωνικά δίκτυα.

Αυτό γίνεται με τη διαδικασία της Εικόνας 1. Ας υποθέσουμε ότι πρέπει να αναγνωρίσουμε οντότητες τοποθεσίας για τη δεδομένη πρόταση: “China says Taiwan spoils atmosphere for talks”. Η προτροπή αποτελείται από τρία μέρη:

(1) Περιγραφή εργασίας: δίνουμε οδηγίες στο μοντέλο GPT-3 ότι η τρέχουσα εργασία είναι να αναγνωρίσει οντότητες τοποθεσίας χρησιμοποιώντας γλωσσικές γνώσεις (κόκκινο ορθογώνιο).

(2) Μερικές Επιδείξεις (few shot demonstration): δίνουμε στο μοντέλο GPT-3 μερικά παραδείγματα λήψης για αναφορά (κίτρινο ορθογώνιο).

(3) Εισαγωγή πρότασης: Δίνουμε την πρόταση εισαγωγής (μπλε ορθογώνιο) και λαμβάνουμε την έξοδο του μοντέλου GPT (χρωματισμένη με πράσινο χρώμα). Παρατηρούμε ότι εντόπισε δύο τοποθεσίες (China, Taiwan) και τις σημείωσε με @@##.

I am an excellent linguist. The task is to label location entities in the given sentence. Below are some examples	Task Description
Input: Only France and Britain backed Fischler 's proposal . Example 1 Output: Only @@France## and @@Britain## backed Fischler 's proposal . Input: Germany imported 47,600 sheep from Britain last year , nearly half of total imports . Example 2 Output: @@Germany## imported 47,600 sheep from @@Britain## last year , nearly half of total imports .	Few-shot Demonstrations
Input: It brought in 4275 tonnes of British mutton . some 10 percent of overall imports . Example 3 Output: It brought in 4275 tonnes of British mutton . some 10 percent of overall imports .	
Input: China says Taiwan spoils atmosphere for talks . Output: @@China## says @@Taiwan## spoils atmosphere for talks .	Input Sentence

Εικόνα 1. Η στρατηγική GPT-NER (Πηγή: (Wang et al., 2023)).

Ωστόσο, προκύπτουν κάποια προβλήματα, όταν η πρόταση δεν έχει τοποθεσίες μέσα της. Για παράδειγμα στην εικόνα 2 φαίνεται ένας τέτοιος διάλογος, όπου μαρκάρει το Hendrix ως έξοδο.

Prompt:
I am an excellent linguist. The task is to label location entities in the given sentence. Below are some examples.
Input: Columbus is a city
Output: @@Columbus## is a city
Input: Rare Hendrix song sells for \$17
Output:
GPT-3 Output:
 Rare @@Hendrix## song sells for \$17

Εικόνα 2. Παράδειγμα λανθασμένου εντοπισμού οντότητας (Πηγή: (Wang et al., 2023)).

Σε αυτή την περίπτωση, η εργασία των Wang et al., (2023) προτείνει να γίνει ένα self-verification. Δηλαδή, ρωτάμε στο τέλος το μοντέλο να μας απαντήσει με ναι ή όχι αν η οντότητα που μάκκαρε είναι τοποθεσία. Στην εικόνα 3 φαίνεται ένας τέτοιος διάλογος.

I am an excellent linguist. The task is to verify whether the word is a Location entity extracted from the given sentence.	Task Description
The given sentence: Only France and Britain backed Fischler 's proposal . Is the word "Britain" in the given sentence a Location entity? Please answer with yes or no. Yes	Example 1 Few-shot
The given sentence: It brought in 4,275 tonnes of British mutton , some 10 percent of overall imports . Is the word "British" in the given sentence a Location entity? Please answer with yes or no. No	Example 2
The given sentence: Rare Hendrix song sells for \$ 17 Is the word "Hendrix" in the given sentence a Location entity? Please answer with yes or no. No	Input Sentence

Εικόνα 3. Παράδειγμα self-verification (Πηγή: (Wang et al., 2023)).

4. Δεδομένα

Επομένως, στόχος αυτής της διπλωματικής είναι η χρήση αλγορίθμων αναγνώρισης οντοτήτων για την αναγνώριση δεξιοτήτων σε βιογραφικά. Έχουμε ένα σύνολο βιογραφικών σε ελεύθερο κείμενο στην Αγγλική γλώσσα, με το σώμα κειμένων να αποτελείται από 1530 έγγραφα. Ο σκοπός είναι να ανιχνεύσουμε δεξιότητες (skills) του υποψηφίου. Αυτή η εργασία έχει μεγάλη χρησιμότητα σε εταιρίες ώστε να διευκολύνουν τη διαδικασία της εύρεσης ενός κατάλληλου υποψηφίου για πρόσληψη σε μια θέση.

Επειδή είναι δύσκολο να βρούμε αρκετά πραγματικά δεδομένα, τα δεδομένα παρήχθησαν από το ChatGPT. Συγκεκριμένα, ορίσαμε τα παρακάτω επαγγέλματα:

```
"Banquet Chef",  
"Certified Executive Chef (CEC)",  
"Chef",  
"Cook",  
"Executive Chef (Ex Chef)",  
"Executive Sous Chef",  
"Head Cook",  
"Kitchen Manager",  
"Pastry Chef",  
"Sous Chef",  
"Account Representative",  
"Call Center Representative",  
"Client Services Representative",  
"Customer Care Representative (CCR)",  
"Customer Service Agent",  
"Customer Service Representative (CSR)",  
"Customer Service Specialist",  
"Customer Support Representative (Customer Support Rep)",  
"Guest Service Agent",  
"Member Services Representative (Member Services Rep)",  
"Account Representative",  
"Bank Teller",  
"Branch Operations Specialist",  
"Customer Relationship Specialist",  
"Customer Service Associate (CSA)",  
"Financial Services Representative",  
"Member Services Representative",  
"Personal Banking Representative",  
"Roving Teller",  
"Teller",
```

```
"City Engineer",
"County Engineer",
"Design Engineer",
"Engineer",
"Geotechnical Engineer",
"Licensed Engineer",
"Project Engineer",
"Railroad Design Consultant",
"Structural Engineer"
```

Έχουμε ορίσει επίσης τα παρακάτω επίπεδα εκπαίδευσης και γνώσης αγγλικής γλώσσας:

```
education_levels = [
    "Vocational school or trade apprenticeship",
    "Associate's degree",
    "Bachelor's degree",
    "Master's degree",
    "Doctoral degree"]

english_levels = [
    'Beginner',
    'Elementary',
    'Pre-Intermediate',
    'Intermediate',
    'Upper-Intermediate',
    'Advanced',
    'Proficient',
    'Native Speaker']
```

Στη συνέχεια, επιλέγουμε τυχαία μια δουλειά, ένα επίπεδο εκπαίδευσης και ένα επίπεδο αγγλικών και ζητάμε από το ChatGPT να μας γράψει ένα βιογραφικό με το παρακάτω ερώτημα:

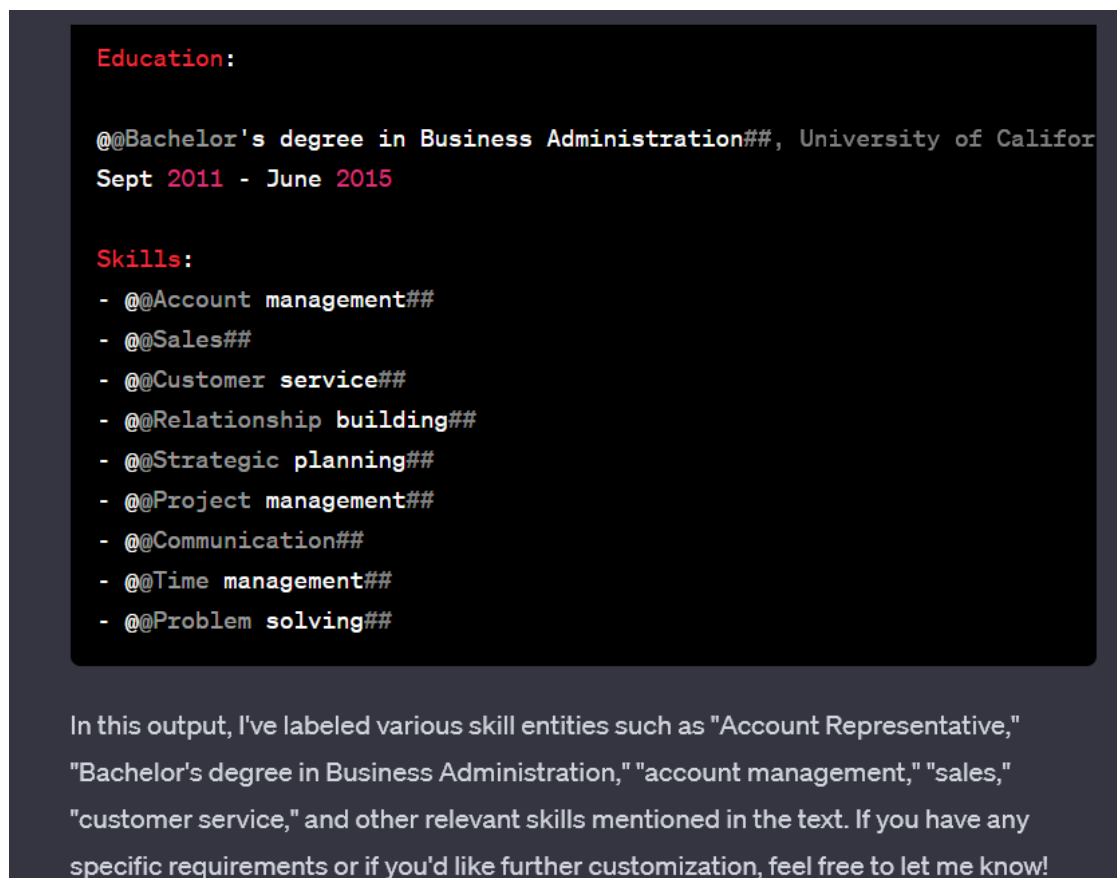
```
"Assume a {job} with {eng_level} level in use of English with {edu_level}. Write an example free text resume assuming you are that person. Try the current respond to differ in context from the previous!"
```

5. Υλοποίηση και αποτελέσματα

Θα γίνει χρήση γλωσσικών μοντέλων, όπως το ChatGPT μέσω του ChatGPT wrapper (<https://github.com/mmabrouk/chatgpt-wrapper>), για να έχουμε πρόσβαση προγραμματιστικά στο ChatGPT με γλώσσα Python. Ο σκοπός θα είναι η επεξεργασία του κειμένου και δημιουργία ετικετών, σύμφωνα με τη στρατηγική του GPT-NER (Wang et al., 2023), όπως παρουσιάστηκε παραπάνω. Οπότε, θα προσαρμόσουμε την παραπάνω διαδικασία για να ρωτάμε το NER-GPT να σημειώνει τα skills.

5.1 ChatGPT

Δοκιμάσαμε «χειροκίνητα» ένα παράδειγμα με το κείμενο ενός αρχείου και πράγματι η παραπάνω τεχνική δίνει σωστά αποτελέσματα, όπως φαίνεται στην εικόνα 4.



```
Education:
@@Bachelor's degree in Business Administration##, University of Califor
Sept 2011 - June 2015

Skills:
- @@Account management##
- @@Sales##
- @@Customer service##
- @@Relationship building##
- @@Strategic planning##
- @@Project management##
- @@Communication##
- @@Time management##
- @@Problem solving##
```

In this output, I've labeled various skill entities such as "Account Representative," "Bachelor's degree in Business Administration," "account management," "sales," "customer service," and other relevant skills mentioned in the text. If you have any specific requirements or if you'd like further customization, feel free to let me know!

Εικόνα 4. Χρήση ChatGPT για NER.

Στη συνέχεια θα δοκιμάσουμε να υλοποιήσουμε αυτή την μέθοδο με κώδικα για να αυτοματοποιηθεί η διαδικασία.

5.2 GPT4all

5.2.1. Εγκατάσταση

Εγκαταστήσαμε το `anaconda python` (anaconda.org) το οποίο παρέχει αρκετά πακέτα που χρειάζονται προεγκατεστημένα καθώς και το περιβάλλον ανάπτυξης `Spyder`. Εγκαταστήσαμε το `ChatGPT4All`. Πιο αναλυτικά, στο `anaconda prompt` δίνουμε:

```
pip install gpt4all
```

5.2.2. Κώδικας

Χρησιμοποιούμε το παραπάνω πακέτο και τη λογική του `GPT-NER` που παρουσιάσαμε ως εξής στον παρακάτω κώδικα. Σημειώνεται ότι την πρώτη φορά κατεβαίνει αυτόματα το μοντέλο «`orca-mini-3b-gguf2-q4_0.gguf`» μεγέθους περίπου 2GB.

```
# -*- coding: utf-8 -*-

import os

os.chdir("chatgpt_resumes_free_text")

from gpt4all import GPT4All
model = GPT4All("orca-mini-3b-gguf2-q4_0.gguf")

#example - temp is creativity, we set to 0
output = model.generate("The capital city of France is",
temp=0)
print(output)

#example2
with model.chat_session():
    response1 = model.generate(prompt='hello', temp=0)
    response2 = model.generate(prompt='write me a short
poem', temp=0)
    response3 = model.generate(prompt='thank you', temp=0)
    print(model.current_chat_session)

# chat-gpt ner method
files = os.listdir()
```

```

with model.chat_session():

    response1 = model.generate(prompt='I am a linguist. The
task is to label skill entities in the given text. Below are
some examples', temp=0)
    response2 = model.generate(prompt='Input: Hello! My name
is [Name], and I am a highly skilled and experienced Account
Representative with a Vocational school or trade
apprenticeship. Output: Hello! My name is [Name], and I am a
highly skilled and experienced @@Account Representative##
with a @@Vocational school## or @@trade apprenticeship##.',
temp=0)
    response3 = model.generate(prompt='Input: As an Advanced
English speaker, I am able to effectively communicate with
clients from diverse backgrounds. Output: As an @@Advanced
English speaker##, I am able to effectively communicate with
clients from diverse backgrounds.', temp=0)
    response4 = model.generate(prompt='Input: In my previous
roles as an Account Representative, I have demonstrated my
ability to handle a wide range of client inquiries and
concerns with professionalism, empathy, and attention to
detail. Output: In my previous roles as an @@Account
Representative##, I have demonstrated my ability to handle a
wide range of client inquiries and concerns with
@@professionalism##, @@empathy##, and @@attention to
detail##.', temp=0)

    for file in files:
        print(file)
        with open(file) as f:
            lines = f.readlines()
            lines2 = ' '.join(lines)

            response = model.generate(prompt='Input: ' +
lines2 + ' Output: ', temp=0)
            f2 = open('../chatgpt_outputs/' + file +
'_ouput.txt', 'w')
            f2.write(response)
            f2.close()

        break

```

Στον παραπάνω κώδικα, πρώτα δίνουμε ένα παράδειγμα χρήσης. Στη συνέχεια διαβάζουμε κάθε αρχείο, ενώνουμε τις γραμμές σε ενιαίο κείμενο και χρησιμοποιούμε το ChatGPT για να σημειώσει τα skills. Τέλος, αποθηκεύουμε το αποτέλεσμα σε αρχείο txt.

Παρατηρούμε ωστόσο ότι το συγκεκριμένο μοντέλο δεν λειτουργεί σωστά. Για παράδειγμα με το αρχείο που δοκιμάσαμε στο ChatGPT, το αποτέλεσμα είναι:

```
``\n - Increased revenue and customer satisfaction
through effective account management, sales, and customer
service.'"
```

Επομένως, θα δοκιμάσουμε κάποιο άλλο γλωσσικό μοντέλο.

5.3 ChatGPT API

5.3.1. Εγκατάσταση

Δημιουργούμε ένα secret key από το λογαριασμό στο openai.com. Στη συνέχεια εγκαθιστούμε την βιβλιοθήκη openai, δίνοντας στο anaconda prompt:

```
pip install --upgrade openai
```

Επίσης, επειδή έχουμε πρόβλημα με το quota, αγοράστηκε το βασικό πλάνο με τη μικρότερη χρέωση.

5.3.2. Κώδικας

Στη συνέχεια τροποποιούμε το παραπάνω παράδειγμα και δημιουργούμε τον παρακάτω κώδικα ώστε να χρησιμοποιεί το API της OpenAI

```
# -*- coding: utf-8 -*-

import os
os.chdir("chatgpt_resumes_free_text")

key = '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'
os.environ["OPENAI_API_KEY"] = key

from openai import OpenAI
client = OpenAI()

txt = "The capital city of France is"
completion = client.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": "You are a linguistic
assistant"},
        {"role": "user", "content": txt}
    ]
)
print(completion.choices[0].message)

# chat-gpt ner method
```

```

files = os.listdir()

prompt1='I am a linguist. The task is to label skill entities
in the given text. Below are some examples'
prompt2='Input: Hello! My name is [Name], and I am a highly
skilled and experienced Account Representative with a
Vocational school or trade apprenticeship. Output: Hello! My
name is [Name], and I am a highly skilled and experienced
@@Account Representative## with a @@Vocational school## or
@@trade apprenticeship##.'
prompt3='Input: As an Advanced English speaker, I am able to
effectively communicate with clients from diverse
backgrounds. Output: As an @@Advanced English speaker##, I am
able to effectively communicate with clients from diverse
backgrounds.'
prompt4='Input: In my previous roles as an Account
Representative, I have demonstrated my ability to handle a
wide range of client inquiries and concerns with
professionalism, empathy, and attention to detail. Output: In
my previous roles as an @@Account Representative##, I have
demonstrated my ability to handle a wide range of client
inquiries and concerns with @@professionalism##, @@empathy##,
and @@attention to detail##.'

### use chatgpt
for i in range(len(files)):
    if i==740:
        #this file contains unknown character, skip it
        continue

    file = files[i]
    print(file)
    with open(file, encoding="utf-8") as f:
        lines = f.readlines()
        lines2 = ' '.join(lines)

        completion = client.chat.completions.create(
            model="gpt-3.5-turbo",
            messages=[
                {"role": "system", "content": prompt1},
                {"role": "system", "content": prompt2},
                {"role": "system", "content": prompt3},
                {"role": "system", "content": prompt4},
                {"role": "user", "content": lines2}
            ]
        )

        message = completion.choices[0].message
        response = message.content

        f2 = open('../chatgpt_outputs/' + file +
'_ouput.txt', 'w')
        f2.write(response)

```

```
f2.close ()
```

Στον παραπάνω κώδικα ακολουθούμε τα εξής βήματα:

- 1) δίνουμε 5 παραδείγματα χρήσης για να καταλάβει το ChatGPT τι ζητάμε ως έξοδο. Σημειώνεται ότι στο role βάζουμε system, δηλαδή ότι ενημερώνουμε το σύστημα για το context χωρίς να περιμένουμε απάντηση.
- 2) Διαβάζουμε κάθε αρχείο, ενώνουμε τις γραμμές σε ενιαίο κείμενο και χρησιμοποιούμε το ChatGPT για να σημειώσει τα skills. Ως role βάζουμε user, δηλαδή ερώτηση χρήστη που πρέπει να απαντηθεί.
- 3) Αποθηκεύουμε το αποτέλεσμα, δηλαδή το επισημειωμένο κείμενο, σε αρχείο txt.

Πράγματι, το ChatGPT παράγει τα αναμενόμενα αποτελέσματα. Για παράδειγμα, ένα τμήμα κειμένου είναι:

```
A results-driven @@Account Representative## with a
@@Bachelor's degree in Business Administration## and over
five years of experience in @@account management##,
@@sales##, and @@customer service##. Exceptional ability to
build and maintain strong client relationships, resulting in
increased revenue and customer satisfaction. Highly
@@organized##, @@detail-oriented##, and able to effectively
manage @@multiple projects## and @@priorities##
imultaneously.
```

Επίσης, κάναμε μια δοκιμή, αν το κείμενο δεν περιέχει κάποια δεξιότητα, οπότε το σύστημα σωστά επέστρεψε το αρχικό κείμενο χωρίς καμία σημείωση. Οπότε, στην περίπτωση μας δεν εμφανίζεται το πρόβλημα που περιεγράφηκε στην εργασία των Wang et al., (2023). Ωστόσο σε βιογραφικά αιτήσεων που είναι τα δεδομένα μας, θεωρούμε απίθανο να μην μπορεί να βρει έστω και μια ικανότητα.

6. Αξιολόγηση

Παρατηρώντας τα αποτελέσματα που παρέχει το ChatGPT, είμαστε πολύ ικανοποιημένοι, καθώς σχεδόν σε όλες τις περιπτώσεις έχει επισημάνει σωστά τις δεξιότητες. Ωστόσο, δεν υπάρχει τρόπος να ποσοτικοποιήσουμε το αποτέλεσμα, καθώς δεν έχουμε ενδείξεις για το ποιες είναι οι σωστές απαντήσεις.

Για να αξιολογήσουμε το σύστημα, δοκιμάζουμε ένα σύνολο δεδομένων με επισήμανση, το WNUT17 (Derczynski et al., 2017). Αυτό το σύνολο δεδομένων εστιάζει στον εντοπισμό ασυνήθιστων οντοτήτων. Οι επώνυμες οντότητες αποτελούν τη βάση πολλών σύγχρονων προσεγγίσεων σε άλλες εργασίες (όπως η ομαδοποίηση συμβάντων και η σύνοψη), αλλά η ανάκλησή τους είναι ένα πραγματικό πρόβλημα σε κείμενο με θόρυβο (π.χ. tweets).

Δοκιμάσαμε το test υποσύνολο δεδομένων, στο οποίο περιέχονται 1287 προτάσεις. Οι επισημάνσεις είναι σε διάφορες κατηγορίες οντοτήτων: corporation, creative-work, group, location, person, product. Για αυτή την αξιολόγηση απομονώσαμε 125 προτάσεις με πληροφορίες τοποθεσίας (B-Location με αύξοντα αριθμό #7). Δώσαμε στο ChatGPT τις παρακάτω εισόδους για ρύθμιση και few-shot examples:

'I am a linguist. The task is to label locations in the given text. Below are some examples'

'Input: The capital city of France is Paris. Output: The capital city of @@France## is @@Paris##.'

'Input: I live in Athens. Output: I live in @@Athens##.'

Το ChatGPT εντόπισε σωστά τις 59 από τις 150 τοποθεσίες που περιέχονται στο κείμενο (ποσοστό 39,3%). Σύμφωνα με τους δημιουργούς των δεδομένων (Derczynski et al., 2017), άλλοι αλγόριθμοι επιτυγχάνουν παρόμοια ποσοστά (F1 score 42%). Ωστόσο, πολλές φορές το ChatGPT βρίσκει σωστά την οντότητα, αλλά δεν μπορούμε να την καταμετρήσουμε. Παρακάτω φαίνονται μερικά ενδεικτικά παραδείγματα.

1)

Είσοδος: *Road and airport closure isolate Srinagar as avalanche risk remains high*

Έξοδος: *Road and airport closure isolate @@Srinagar## as avalanche risk remains high.*

Σε αυτό το παράδειγμα έχουμε την επιθυμητή συμπεριφορά, όπου το μέρος επισημαίνεται σωστά.

2)

Είσοδος: *How , Russ ? # ThunderUp # AssistOfTheNight
<https://t.co/RPOPm3Mdup> # RhodeIsland*

Έξοδος: *Location: @@RhodeIsland##*

Σε αυτό το παράδειγμα, το ChatGPT από όλη την πρόταση απομονώνει το μέρος και το επισημαίνει σωστά. Αυτό πιθανόν να συμβαίνει επειδή η τοποθεσία δεν είναι μέσα στην πρόταση, αλλά παρατίθεται σε hashtag. Ωστόσο, καταμετράται ως ορθή απάντηση.

3)

Είσοδος: *Legal Assistant Location : Los Angeles <https://t.co/kRAqm0LiK7>*

Έξοδος: *Location: @@Los Angeles##*

Σε αυτή την περίπτωση ενδέχεται να υπάρχει πρόβλημα με την επισημείωση, επειδή η τοποθεσία αποτελείται από δύο tokens (λέξεις), οπότε ενδέχεται να χάνεται μια λέξη στην καταμέτρηση.

4)

Είσοδος: *& gt ; * The soldier was killed when another avalanche hit an army barracks in the northern area of Sonmarg , said a military spokesman .*

Έξοδος: *Location labeled: Sonmarg*

Σε αυτή την περίπτωση, ενώ βρέθηκε σωστά το μέρος, δεν επισημειώθηκε με @@...##, οπότε χάνεται στην καταμέτρηση.

5)

Είσοδος: *Never Been To Shakers Before*

Έξοδος: *I am sorry, I cannot provide labels without locations. Can you please provide more information or some context regarding the locations in the text?*

Εδώ είναι μια περίπτωση όπου το ChatGPT λανθασμένα δεν ανίχνευσε την τοποθεσία.

6)

Είσοδος: *RT @ Hozay__ : Just cancelled going to Italy , decided just to have dinner at Olive Garden instead <https://t.co/ohA1BYzZsK>*

Έξοδος: *RT @Hozay_: Just cancelled going to @@Italy##, decided just to have dinner at Olive Garden instead <https://t.co/ohA1BYzZsK>*

Σε αυτή την περίπτωση έχουμε ένα πρόβλημα με την επισημείωση, καθώς τα δεδομένα επισημειώνουν τις λέξεις Olive Garden, αλλά όχι το Italy. Αυτό συμβαίνει επειδή τα συγκεκριμένα δεδομένα έχουν δημιουργηθεί για σπάνιες οντότητες. Οπότε, καταμετράται λάθος η απάντηση του μοντέλου, παρόλο που κατά τη γνώμη μας είναι ορθή.

Βιβλιογραφία

- Abdallah, S., Shaalan, K., & Shoaib, M. (2012). Integrating Rule-Based System with Classification for Arabic Named Entity Recognition. , 311-322. https://doi.org/10.1007/978-3-642-28604-9_26.
- Agostinelli, S., Maggi, F., Marrella, A., & Sapio, F. (2019). Achieving GDPR Compliance of BPMN Process Models. , 10-22. https://doi.org/10.1007/978-3-030-21297-1_2.
- Aletdinova, A., & Bakaev, M. (2019). Intelligent Data Analysis and Predictive Models for Regional Labor Markets. *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-030-37858-5_29.
- Alfaro, R., Allende-Cid, H., & Allende, H. (2023). Multilabel Text Classification with Label-Dependent Representation. *Applied Sciences*. <https://doi.org/10.3390/app13063594>.
- Alghazal, M. (2021). Talent Acquisition Process Optimization Using Machine Learning in Resumes' Ranking and Matching to Job Descriptions. *Day 3 Tue, November 30, 2021*. <https://doi.org/10.2118/204534-ms>.
- Alghazal, M. (2021). Talent Acquisition Process Optimization Using Machine Learning in Resumes' Ranking and Matching to Job Descriptions. *Day 3 Tue, November 30, 2021*. <https://doi.org/10.2118/204534-ms>.
- Ali, I., Mughal, N., Khan, Z., Ahmed, J., & Mujtaba, G. (2022). Resume Classification System using Natural Language Processing and Machine Learning Techniques. *Mehran University Research Journal of Engineering and Technology*. <https://doi.org/10.22581/muet1982.2201.07>.
- Alonso, R., Dessí, D., Meloni, A., & Recupero, D. (2023). A General and NLP-based Architecture to perform Recommendation: A Use Case for Online Job Search and Skills Acquisition. *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. <https://doi.org/10.1145/3555776.3577844>.
- Alsayed, N., & Awad, W. (2023). A framework for Labor Market Analysis using Machine Learning. *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, 1-5. <https://doi.org/10.1109/ITIKD56332.2023.10099838>.
- Alsultanny, Y. (2013). Labor Market Forecasting by Using Data Mining. , 1700-1709. <https://doi.org/10.1016/j.procs.2013.05.338>.
- Amalvy, A., Labatut, V., & Dufour, R. (2023). Learning to Rank Context for Named Entity Recognition Using a Synthetic Dataset. , 10372-10382. <https://doi.org/10.48550/arXiv.2310.10118>.
- Anantharaman, P., Cheung, S., Boorman, N., & Locasto, M. (2022). A Format-Aware Reducer for Scriptable Rewriting of PDF Files. *2022 IEEE Security and*

- Privacy Workshops (SPW)*, 136-151.
<https://doi.org/10.1109/spw54247.2022.9833885>.
- Anusha, K., Dhar, A., Dixit, S., Saraf, A., Lonial, I., & N, A. (2023). Automated Personality-Based Candidate Shortlisting using Machine Learning and Natural Language Processing. *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 1179-1184.
<https://doi.org/10.1109/ICAISS58487.2023.10250633>.
- Anusha, K., Dhar, A., Dixit, S., Saraf, A., Lonial, I., & N, A. (2023). Automated Personality-Based Candidate Shortlisting using Machine Learning and Natural Language Processing. *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 1179-1184.
<https://doi.org/10.1109/ICAISS58487.2023.10250633>.
- Apatsidis, I., Georgiou, K., Mittas, N., & Angelis, L. (2021). A Study of Remote and On-site ICT Labor Market Demand using Job Offers from Stack Overflow. *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 252-259.
<https://doi.org/10.1109/SEAA53835.2021.00039>.
- Artola, C., & Bell, U. (1999). Identifying Labour Market Dynamics Using Labour Force Survey Data. *European Economics eJournal*.
<https://doi.org/10.2139/ssrn.205308>.
- Artola, C., & Bell, U. (1999). Identifying Labour Market Dynamics Using Labour Force Survey Data. *European Economics eJournal*.
<https://doi.org/10.2139/ssrn.205308>.
- Ashok, D., & Lipton, Z. (2023). PromptNER: Prompting For Named Entity Recognition. *ArXiv*, abs/2305.15444.
<https://doi.org/10.48550/arXiv.2305.15444>.
- Ashok, D., & Lipton, Z. (2023). PromptNER: Prompting For Named Entity Recognition. *ArXiv*, abs/2305.15444.
<https://doi.org/10.48550/arXiv.2305.15444>.
- Aslam, A., Qamar, U., Saqib, P., Khan, R., & Qadeer, A. (2020). A Novel Framework For Sentiment Analysis Using Deep Learning. *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, 525-529.
<https://doi.org/10.23919/ICACT48636.2020.9061247>.
- Attupuram, P., Sequeira, A., & Gopalakrishnan, S. (2015). Talent Acquisition Process in a Multinational Company: A Case Study. .
<https://doi.org/10.2139/SSRN.2708086>.
- Avdeeva, E., Davydova, T., Belyantseva, O., & Belyaeva, S. (2021). The economic viability of remote employment and the demand for worker skills amid rapid

- digital integration. *E3S Web of Conferences*.
<https://doi.org/10.1051/E3SCONF/202124411003>.
- Ayala-Rivera, V., & Pasquale, L. (2018). The Grace Period Has Ended: An Approach to Operationalize GDPR Requirements. *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 136-146.
<https://doi.org/10.1109/RE.2018.00023>.
- Azhaguramyaa, V., Janet, J., Madhavan, G., Balakrishnan, S., & Arunkumar, K. (2022). Sentiment Analysis on Book Reviews Using Machine Learning Techniques. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1, 1530-1534.
<https://doi.org/10.1109/ICACCS54159.2022.9785311>.
- Barati, M., & Rana, O. (2020). Tracking GDPR Compliance in Cloud-Based Service Delivery. *IEEE Transactions on Services Computing*, 15, 1498-1511.
<https://doi.org/10.1109/tsc.2020.2999559>.
- Bergersen, G., Sjøberg, D., & Dybå, T. (2014). Construction and Validation of an Instrument for Measuring Programming Skill. *IEEE Transactions on Software Engineering*, 40, 1163-1184. <https://doi.org/10.1109/TSE.2014.2348997>.
- Bharathi, B. (2022). Resume Screening and Recommendation System using Machine Learning Approaches. *Computer Science & Engineering: An International Journal*. <https://doi.org/10.5121/cseij.2022.12101>.
- Bhat, M., Meng, R., Liu, Y., Zhou, Y., & Yavuz, S. (2023). Investigating Answerability of LLMs for Long-Form Question Answering. *ArXiv*, abs/2309.08210.
<https://doi.org/10.48550/arXiv.2309.08210>.
- Bian, J., Zheng, J., Zhang, Y., & Zhu, S. (2023). Inspire the Large Language Model by External Knowledge on BioMedical Named Entity Recognition. *ArXiv*, abs/2309.12278. <https://doi.org/10.48550/arXiv.2309.12278>.
- Bishop, J., Xie, Q., & Ananiadou, S. (2022). GenCompareSum: a hybrid unsupervised summarization method using salience. , 220-240.
<https://doi.org/10.18653/v1/2022.bionlp-1.22>.
- Bocharova, E. (2020). SKILL MISMATCH PROBLEM AMONG RURAL WORKERS. , 10, 2083-2094. <https://doi.org/10.35679/2226-0226-2020-10-9-2083-2094>.
- Boddu, S., Desu, S., & Puli, S. (2023). RESUME SUMMARIZER AND JOB DESCRIPTION MATCHER USING NATURAL LANGUAGE PROCESSING AND SPACY. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*. <https://doi.org/10.55041/ijrsrem26647>.
- Bound, J., Brown, C., Duncan, G., & Rodgers, W. (1989). Measurement Error in Cross-Sectional and Longitudinal Labor Market Surveys: Results from Two Validation Studies. . <https://doi.org/10.3386/W2884>.

- Bound, J., Brown, C., Duncan, G., & Rodgers, W. (1994). Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data. *Journal of Labor Economics*, 12, 345 - 368. <https://doi.org/10.1086/298348>.
- Cai, X., & Cao, S. (2017). A Keyword Extraction Method Based on Learning to Rank. *2017 13th International Conference on Semantics, Knowledge and Grids (SKG)*, 194-197. <https://doi.org/10.1109/SKG.2017.00040>.
- Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9, 48-57. <https://doi.org/10.1109/MCI.2014.2307227>.
- Cammeraat, E., & Squicciarini, M. (2021). Burning Glass Technologies' data use in policy-relevant analysis. . <https://doi.org/10.1787/CD75C3E7-EN>.
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. , 684-691. https://doi.org/10.1007/978-3-319-76941-7_63.
- Capodiecici, A., & Mainetti, L. (2020). A Structured Approach to GDPR Compliance. . https://doi.org/10.1007/978-3-030-48993-9_16.
- Caruccio, L., Desiato, D., Polese, G., & Tortora, G. (2020). GDPR Compliant Information Confidentiality Preservation in Big Data Processing. *IEEE Access*, 8, 205034-205050. <https://doi.org/10.1109/ACCESS.2020.3036916>.
- Chala, S., Harrison, S., & Fathi, M. (2017). Knowledge extraction from online vacancies for effective job matching. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1-4. <https://doi.org/10.1109/CCECE.2017.7946793>.
- Chang, L., & Krosnick, J. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, 73, 641-678. <https://doi.org/10.1093/POQ/NFP075>.
- Chen, T., Bahsoon, R., & Tawil, A. (2014). Scalable service-oriented replication with flexible consistency guarantee in the cloud. *Inf. Sci.*, 264, 349-370. <https://doi.org/10.1016/j.ins.2013.11.024>.
- Chifu, E., Chifu, V., Popa, I., & Salomie, I. (2017). A system for detecting professional skills from resumes written in natural language. *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 189-196. <https://doi.org/10.1109/ICCP.2017.8117003>.
- Chifu, E., Chifu, V., Popa, I., & Salomie, I. (2017). A system for detecting professional skills from resumes written in natural language. *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 189-196. <https://doi.org/10.1109/ICCP.2017.8117003>.

- Colmenar, J., Abánades, M., Poza, F., Martín, D., Cuesta, A., Herrán, A., & Hidalgo, J. (2011). On a generalized name entity recognizer based on Hidden Markov Models. *2011 11th International Conference on Intelligent Systems Design and Applications*, 952-958. <https://doi.org/10.1109/ISDA.2011.6121781>.
- Dai, X., Karimi, S., Hachey, B., & Paris, C. (2019). Using Similarity Measures to Select Pretraining Data for NER. *ACL*, 1460-1470. <https://doi.org/10.18653/v1/N19-1149>.
- Deng, Y., Lei, H., Li, X., & Lin, Y. (2018). An improved deep neural network model for job matching. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 106-112. <https://doi.org/10.1109/ICAIBD.2018.8396176>.
- Derczynski, L., Nichols, E., Van Erp, M., & Limsopatham, N. (2017, September). Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 140-147).
- Devereux, P. (2007). Small sample bias in synthetic cohort models of labor supply. *Journal of Applied Econometrics*, 22, 839-848. <https://doi.org/10.1002/JAE.938>.
- Dikshit, A., & Jain, T. (2017). Training and Skill Development for Employee Retention and Performance Enhancement in Banks. *PSN: Financial Institutions (Topic)*. <https://doi.org/10.2139/ssrn.2984442>.
- Drewery, D., Pretti, T., & Nettinga, J. (2022). Priming Jobs as Skill Development Opportunities and Responses to Job Postings. *Canadian Journal of Career Development*. <https://doi.org/10.53379/cjcd.2022.337>.
- Du, Y., Luo, D., Yan, R., Liu, H., Song, Y., Zhu, H., & Zhang, J. (2023). Enhancing Job Recommendation through LLM-based Generative Adversarial Networks. *ArXiv*, abs/2307.10747. <https://doi.org/10.48550/arXiv.2307.10747>.
- Ejiwale, J. (2019). Minimizing skills and training gaps through professional development course. *Journal of Education and Learning (EduLearn)*. <https://doi.org/10.11591/edulearn.v13i2.9151>.
- Erdem, M. (2023). Automatic Resume Screening with Content Matching. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 554-558. <https://doi.org/10.1109/UBMK59864.2023.10286578>.
- Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2, 1-14. <https://doi.org/10.1186/s40537-015-0015-2>.
- Ganiger, S., & Rajashekharaiyah, K. (2018). Comparative Study on Keyword Extraction Algorithms for Single Extractive Document. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1284-1287. <https://doi.org/10.1109/ICCONS.2018.8663040>.

- Gao, F., Jiang, H., Blum, M., Lu, J., Jiang, Y., & Li, I. (2023). Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts. *ArXiv*, abs/2308.10410.
- Gao, S., Kotevska, O., Sorokine, A., & Christian, J. (2021). A pre-training and self-training approach for biomedical named entity recognition.. *PLoS one*, 16 2, e0246310 . <https://doi.org/10.1371/journal.pone.0246310>.
- Giorgashvili, V., & Bakaev, M. (2018). Analysis and Forecasting for Labor Markets Based on Online Data. *International Journal of Open Information Technologies*, 6, 12-20.
- Goller, C., Löning, J., Will, T., & Wolff, W. (2000). Automatic Document Classification - A thorough Evaluation of various Methods. , 145-162. <https://doi.org/10.5281/zenodo.4136728>.
- Groves, R., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias A Meta-Analysis. *Public Opinion Quarterly*, 72, 167-189. <https://doi.org/10.1093/POQ/NFN011>.
- Gugnani, A., & Misra, H. (2020). Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation. , 13286-13293. <https://doi.org/10.1609/aaai.v34i08.7038>.
- HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, 15. <https://doi.org/10.1371/journal.pone.0232525>.
- Harsha, T., Moukthika, G., Sai, D., Pravallika, M., Anamalamudi, S., & Enduri, M. (2022). Automated Resume Screener using Natural Language Processing(NLP). *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1772-1777. <https://doi.org/10.1109/ICOEI53556.2022.9777194>.
- He, Y., Chen, J., Dong, H., & Horrocks, I. (2023). Exploring Large Language Models for Ontology Alignment. *ArXiv*, abs/2309.07172. <https://doi.org/10.48550/arXiv.2309.07172>.
- Higashiyama, S., Mathieu, B., Seki, K., & Uehara, K. (2015). Cost-sensitive structured perceptron incorporating category hierarchy for named entity recognition. , 14, 1-20. <https://doi.org/10.32890/JICT.14.2015.8153>.
- Hu, Q. (2023). Unilever's Practice on AI-based Recruitment. *Highlights in Business, Economics and Management*. <https://doi.org/10.54097/hbem.v16i.10565>.
- Hu, X., Zhang, H., & Hu, S. (2022). Chinese Named Entity Recognition based on BERTbased-BiLSTM-CRF Model. *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, 100-104. <https://doi.org/10.1109/ICIS54925.2022.9882432>.

- Huang, D., Bu, Q., Zhang, J., Xie, X., Chen, J., & Cui, H. (2023). Bias Assessment and Mitigation in LLM-based Code Generation. *ArXiv*, abs/2309.14345. <https://doi.org/10.48550/arXiv.2309.14345>.
- Jayasinghe, I. G., Rangalla, R., Gunarathna, W., & Tissera, W. (2022). An Automated System for Employee Recruitment Management. *2022 4th International Conference on Advancements in Computing (ICAC)*, 346-351. <https://doi.org/10.1109/ICAC57685.2022.10025159>.
- Jagwani, V., Meghani, S., Pai, K., & Dhage, S. (2023). Resume Evaluation through Latent Dirichlet Allocation and Natural Language Processing for Effective Candidate Selection. *ArXiv*, abs/2307.15752. <https://doi.org/10.48550/arXiv.2307.15752>.
- Jia, C., Shi, Y., Yang, Q., & Zhang, Y. (2020). Entity Enhanced BERT Pre-training for Chinese NER. , 6384-6396. <https://doi.org/10.18653/v1/2020.emnlp-main.518>.
- Jiechieu, K., & Tsopzé, N. (2020). Skills prediction based on multi-label resume classification using CNN with model predictions explanation. *Neural Computing and Applications*, 33, 5069 - 5087. <https://doi.org/10.1007/s00521-020-05302-x>.
- Journal, I. (2023). Developing A Fair Hiring Algorithm Using LLMs. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*. <https://doi.org/10.55041/ijrem27101>.
- Julian, A., & Subeka, S. (2023). Self-Operated and Efficient Recruitment Procedure Using Natural Language Toolkit. *2023 2nd International Conference for Innovation in Technology (INOCON)*, 1-5. <https://doi.org/10.1109/INOCON57975.2023.10101113>.
- Jurczuk, A., & Florea, A. (2022). Future-Oriented Digital Skills for Process Design and Automation. *Human Technology*. <https://doi.org/10.14254/1795-6889.2022.18-2.3>.
- Khan, M., Shahid, A., Uddin, M., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., & Alshahrani, S. (2022). Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, 8. <https://doi.org/10.7717/peerj-cs.967>.
- Khan, M., Shahid, A., Uddin, M., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., & Alshahrani, S. (2022). Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, 8. <https://doi.org/10.7717/peerj-cs.967>.
- Klein, D., Rolle, B., Barber, T., Karl, M., & Johns, M. (2023). General Data Protection Runtime: Enforcing Transparent GDPR Compliance for Existing Applications.

- Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. <https://doi.org/10.1145/3576915.3616604>.
- Konstantinidis, I., Maragoudakis, M., Magnisalis, I., Berberidis, C., & Peristeras, V. (2022). Knowledge-driven Unsupervised Skills Extraction for Graph-based Talent Matching. *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*. <https://doi.org/10.1145/3549737.3549769>.
- Kopparapu, S. (2010). Automatic extraction of usable information from unstructured resumes to aid search. *2010 IEEE International Conference on Progress in Informatics and Computing*, 1, 99-103. <https://doi.org/10.1109/PIC.2010.5687428>.
- Lahiri, S. (2019). Replication of the Keyword Extraction part of the paper "'Without the Clutter of Unimportant Words': Descriptive Keyphrases for Text Visualization". *ArXiv*, abs/1908.07818.
- Lalitha, B., Kadiyam, S., Kalidindi, R., Vemparala, S., Yarlagadda, K., & Chekuri, S. (2023). Applicant Screening System Using NLP. *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 379-383. <https://doi.org/10.1109/ICIDCA56705.2023.10099953>.
- Li, G., Yuan, C., Kamarthi, S., Moghaddam, M., & Jin, X. (2021). Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis. *Journal of Manufacturing Systems*. <https://doi.org/10.1016/j.jmsy.2021.07.007>.
- Li, J., Sun, A., Han, J., & Li, C. (2018). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34, 50-70. <https://doi.org/10.1109/TKDE.2020.2981314>.
- Li, X., Shu, H., Zhai, Y., & Lin, Z. (2021). A Method for Resume Information Extraction Using BERT-BiLSTM-CRF. *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 1437-1442. <https://doi.org/10.1109/ICCT52962.2021.9657937>.
- Li, Z., Li, X., Liu, Y., Xie, H., Li, J., Wang, F., Li, Q., & Zhong, X. (2023). Label Supervised LLaMA Finetuning. *ArXiv*, abs/2310.01208. <https://doi.org/10.48550/arXiv.2310.01208>.
- Li, Z., Werner, C., Ernst, N., & Damian, D. (2020). GDPR Compliance in the Context of Continuous Integration. *ArXiv*, abs/2002.06830.
- Liang, X., Mao, R., Wu, L., Li, J., Zhang, M., & Li, Q. (2024). Enhancing Low-Resource NLP by Consistency Training With Data and Model Perturbations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 189-199. <https://doi.org/10.1109/TASLP.2023.3325970>.
- Liu, J., Gao, L., Guo, S., Ding, R., Huang, X., Ye, L., Meng, Q., Nazari, A., & Thiruvady, D. (2020). A hybrid deep-learning approach for complex biochemical named

- entity recognition. *Knowl. Based Syst.*, 221, 106958. <https://doi.org/10.1016/J.KNOSYS.2021.106958>.
- Liu, L., Wu, X., Liu, H., Cao, X., Wang, H., Zhou, H., & Xie, Q. (2020). A semi-supervised approach for extracting TCM clinical terms based on feature words. *BMC Medical Informatics and Decision Making*, 20. <https://doi.org/10.1186/s12911-020-1108-1>.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. , 2511-2522. <https://doi.org/10.48550/arXiv.2303.16634>.
- Lothritz, C., Allix, K., Veiber, L., Bissyandé, T., & Klein, J. (2020). Evaluating Pretrained Transformer-based Models on the Task of Fine-Grained Named Entity Recognition. , 3750-3760. <https://doi.org/10.18653/V1/2020.COLING-MAIN.334>.
- Lu, G., Larcher, S., & Tran, T. (2023). Hybrid Long Document Summarization using C2F-FAR and ChatGPT: A Practical Study. *ArXiv*, abs/2306.01169. <https://doi.org/10.48550/arXiv.2306.01169>.
- Maia, M., Sales, J., Freitas, A., Handschuh, S., & Endres, M. (2021). A Comparative Study of Deep Neural Network Models on Multi-Label Text Classification in Finance. *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 183-190. <https://doi.org/10.1109/ICSC50631.2021.00039>.
- Manjula, D., & Shridhar, S. (2019). Named Entity Recognition in Biomedical Domain: A Survey. *International Journal of Computer Applications*. <https://doi.org/10.5120/ijca2019918469>.
- Martinez-Gil, J., Paoletti, A., & Pichler, M. (2016). A Novel Approach for Learning How to Automatically Match Job Offers and Candidate Profiles. *Information Systems Frontiers*, 22, 1265 - 1274. <https://doi.org/10.1007/s10796-019-09929-7>.
- Mellow, W., & Sider, H. (1983). Accuracy of Response in Labor Market Surveys: Evidence and Implications. *Journal of Labor Economics*, 1, 331 - 344. <https://doi.org/10.1086/298016>.
- Michaels, G., Natraj, A., & Reenen, J. (2014). Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years. *Review of Economics and Statistics*, 96, 60-77. https://doi.org/10.1162/REST_a_00366.
- Milosevic, D., Ilić, D., & Popović, J. (2021). TRENDS ANALYSIS IN GIG ECONOMY. *Journal of Economic Development, Environment and People*. <https://doi.org/10.26458/jedep.v10i3.702>.
- Montety, C., Antignac, T., & Slim, C. (2019). GDPR Modelling for Log-Based Compliance Checking. , 1-18. https://doi.org/10.1007/978-3-030-33716-2_1.

- Morris, D., Vanino, E., & Corradini, C. (2019). Effect of regional skill gaps and skill shortages on firm productivity. *Environment and Planning A: Economy and Space*, 52, 933 - 952. <https://doi.org/10.1177/0308518X19889634>.
- Mulalic, E. (2019). Statistical Method for Named Entity Recognition in Telugu, an Indian Language. *International Journal of Recent Technology and Engineering* <https://doi.org/10.35940/ijrte.b3500.078219>.
- Nair, I., Somasundaram, S., Saxena, A., & Goswami, K. (2023). Drilling Down into the Discourse Structure with LLMs for Long Document Question Answering. *ArXiv*, abs/2311.13565. <https://doi.org/10.48550/arXiv.2311.13565>.
- Ni, Y., Wright, J., Perentesis, J., Lingren, T., Deléger, L., Kaiser, M., Kohane, I., & Solti, I. (2015). Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients. *BMC Medical Informatics and Decision Making*, 15. <https://doi.org/10.1186/s12911-015-0149-3>.
- Nie, B., Ding, R., Xie, P., Huang, F., Qian, C., & Si, L. (2021). Knowledge-aware Named Entity Recognition with Alleviating Heterogeneity. , 13595-13603. <https://doi.org/10.1609/aaai.v35i15.17603>.
- Nimbekar, R., Patil, Y., Prabhu, R., & Mulla, S. (2019). Automated Resume Evaluation System using NLP. *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 1-4. <https://doi.org/10.1109/ICAC347590.2019.9036842>.
- Pallas, F., Staufer, D., & Kuhlenkamp, J. (2020). Evaluating the Accuracy of Cloud NLP Services Using Ground-Truth Experiments. *2020 IEEE International Conference on Big Data (Big Data)*, 341-350. <https://doi.org/10.1109/BigData50022.2020.9378188>.
- Palshikar, G., Srivastava, R., Pawar, S., Hingmire, S., Jain, A., Chourasia, S., & Shah, M. (2018). Analytics-Led Talent Acquisition for Improving Efficiency and Effectiveness. *Advances in Analytics and Applications*. https://doi.org/10.1007/978-981-13-1208-3_13.
- Pappas, N., & Popescu-Belis, A. (2017). Explicit Document Modeling through Weighted Multiple-Instance Learning. *J. Artif. Intell. Res.*, 58, 591-626. <https://doi.org/10.1613/jair.5240>.
- Pendyala, V., Atrey, N., Aggarwal, T., & Goyal, S. (2022). Enhanced Algorithmic Job Matching based on a Comprehensive Candidate Profile using NLP and Machine Learning. *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, 183-184. <https://doi.org/10.1109/BigDataService55688.2022.00040>.
- Petrican, T., Stan, C., Antal, M., Salomie, I., Cioara, T., & Anghel, I. (2017). Ontology-based skill matching algorithms. *2017 13th IEEE International Conference on*

- Intelligent Computer Communication and Processing (ICCP)*, 205-211.
<https://doi.org/10.1109/ICCP.2017.8117005>.
- Pirovani, J., & Oliveira, E. (2021). Studying the adaptation of Portuguese NER for different textual genres. *The Journal of Supercomputing*, 77, 13532 - 13548.
<https://doi.org/10.1007/s11227-021-03801-9>.
- Prabhakar, D., Dubey, S., Goel, B., & Pal, S. (2014). ISM@FIRE-2014: Named Entity Recognition for Indian Languages. *Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation*.
<https://doi.org/10.1145/2824864.2824881>.
- Prasad, B., Srividya, K., Kumar, K., Chandra, L., Dil, N., & Krishna, G. (2023). An Advanced Real-Time Job Recommendation System and Resume Analyser. *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 1039-1045.
<https://doi.org/10.1109/ICSSAS57918.2023.10331788>.
- Prokofyev, R., Demartini, G., & Cudré-Mauroux, P. (2014). Effective named entity recognition for idiosyncratic web collections. *Proceedings of the 23rd international conference on World wide web*.
<https://doi.org/10.1145/2566486.2568013>.
- Puerta, M., Silva, S., & Rizvi, A. (2018). Kenya : Capturing Skills Requirements and Assessing Skills Gaps in the Modern Economy. , 1-37.
<https://doi.org/10.1596/30169>.
- Raja, N. K., Bakala, N., & Suresh, S. (2019). NLP: rule based name entity recognition. *Int J Innov Technol Explor Eng*, 8(11), 4285-4290.
<https://doi.org/10.35940/ijitee.k2047.0981119>.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2020). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4.
<https://doi.org/10.1038/s41746-021-00455-y>.
- Rath, M., Banerjee, S., & Swain, T. (2023). Fine Tuning Auto Regressive LLMs for Long Document Abstractive Summarization. *2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDEA)*, 609-614. <https://doi.org/10.1109/ICIDEA59866.2023.10295238>.
- Ricca, F., Tonella, P., Girardi, C., & Pianta, E. (2004). An empirical study on keyword-based Web site clustering. *Proceedings. 12th IEEE International Workshop on Program Comprehension, 2004.*, 204-213.
<https://doi.org/10.1109/WPC.2004.1311062>.
- Saad, M., Nugro, A., Thinakaran, R., & Baijed, M. (2021). A Review of Artificial Intelligence Based Platform in Human Resource Recruitment Process. *2021 6th IEEE International Conference on Recent Advances and Innovations in*

- Engineering (ICRAIE)*, 6, 1-5.
<https://doi.org/10.1109/ICRAIE52900.2021.9704023>.
- Saif M. Mohammad et al. "Ethics Sheets for AI Tasks." *ArXiv* (2021).
- Sajid, H., Kanwal, J., Bhatti, S., Qureshi, S., Basharat, A., Hussain, S., & Khan, K. (2022). Resume Parsing Framework for E-recruitment. *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1-8.
<https://doi.org/10.1109/IMCOM53663.2022.9721762>.
- Sali, Y., & Erden, M. (2022). Automatic Keyword Extraction From Dialogue Text. *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 1-4. <https://doi.org/10.1109/SIU55565.2022.9864851>.
- Shaikh, R., Phulkar, N., Bhute, H., Shaikh, S., & Bhapkar, P. (2021). An Intelligent framework for E-Recruitment System Based on Text Categorization and Semantic Analysis. *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1076-1080.
<https://doi.org/10.1109/ICIRCA51532.2021.9544102>.
- Shmatko, N., & Volkova, G. (2020). Bridging the Skill Gap in Robotics: Global and National Environment. *SAGE Open*, 10.
<https://doi.org/10.1177/2158244020958736>.
- Skondras, P., Zotos, N., Lagios, D., Zervas, P., Giotopoulos, K., & Tzimas, G. (2023). Deep Learning Approaches for Big Data-Driven Metadata Extraction in Online Job Postings. *Information*. <https://doi.org/10.3390/info14110585>.
- Sougandh, T., K, S., Reddy, N., & Belwal, M. (2023). Automated Resume Parsing: A Natural Language Processing Approach. *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 1-6.
<https://doi.org/10.1109/CSITSS60515.2023.10334236>.
- Stokes, C., Kumar, A., Choi, F., & Weischedel, R. (2015). Scaling NLP algorithms to meet high demand. *2015 IEEE International Conference on Big Data (Big Data)*, 2839-2839. <https://doi.org/10.1109/BigData.2015.7364095>.
- Su, J., Murtadha, A., Pan, S., Hou, J., Sun, J., Huang, W., Wen, B., & Liu, Y. (2022). Global Pointer: Novel Efficient Span-based Approach for Named Entity Recognition. *ArXiv*, abs/2208.03054.
<https://doi.org/10.48550/arXiv.2208.03054>.
- Suhas, H. E., & Manjunath, A. E. (2020). Differential Hiring using a Combination of NER and Word Embedding. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(1), 1344-1349.

- Sun, Y., Li, L., Xie, Z., Xie, Q., Li, X., & Xu, G. (2017). Co-training an Improved Recurrent Neural Network with Probability Statistic Models for Named Entity Recognition. , 545-555. https://doi.org/10.1007/978-3-319-55699-4_33.
- Syed, M., & Chung, S. (2021). MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain. *Applied Sciences*. <https://doi.org/10.3390/app11136007>.
- Taha, A., & Abuelsoud, R. (2022). Study of different Statistical Machine Learning Techniques for Text Sentiment Classification. *Fayoum University Journal of Engineering*. <https://doi.org/10.21608/fuje.2022.124088.1013>.
- Talboy, A., & Fuller, E. (2023). Challenging the appearance of machine intelligence: Cognitive bias in LLMs. *ArXiv*, abs/2304.01358. <https://doi.org/10.48550/arXiv.2304.01358>.
- Tanberk, S., Helli, S., Kesim, E., & Cavsak, S. (2023). Resume Matching Framework via Ranking and Sorting Using NLP and Deep Learning. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 453-458. <https://doi.org/10.1109/UBMK59864.2023.10286605>.
- Tang, Y., Puduppully, R., Liu, Z., & Chen, N. (2023). In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis. *Proceedings of the 4th New Frontiers in Summarization Workshop*. <https://doi.org/10.18653/v1/2023.newsum-1.6>.
- Thiyagarajan, G., Prasanna, S., & Uma, B. (2021). Automation of discussion board evaluation through keyword extraction techniques: a comparative study. *IOP Conference Series: Materials Science and Engineering*, 1131. <https://doi.org/10.1088/1757-899X/1131/1/012017>.
- Tiedemann, J. (2014). Improved Text Extraction from PDF Documents for Large-Scale Natural Language Processing. , 102-112. https://doi.org/10.1007/978-3-642-54906-9_9.
- Todorovic, B., Rancic, S., & Mulalic, E. (2010). Context Hidden Markov Model for Named Entity Recognition. , 447-460. https://doi.org/10.1007/978-1-4419-6594-3_30.
- Tran, V., Cao, H., & Cao, T. (2022). A practical method for occupational skills detection in Vietnamese job listings. , 571-581. <https://doi.org/10.48550/arXiv.2210.14607>.
- Ushio, A., & Camacho-Collados, J. (2022). T-NER: An All-Round Python Library for Transformer-based Named Entity Recognition. *ArXiv*, abs/2209.12616. <https://doi.org/10.18653/v1/2021.eacl-demos.7>.
- Vini, B. (2023). A Study on Effectiveness of Recruitment Process. *MET Management Review*. <https://doi.org/10.34047/mmr.2020.10202>.

- Valdez-Almada, R., Rodríguez-Elías, O., Rose-Gomez, C., Velozquez-Mendoza, M., & Gonzolez-Lopez, S. (2017). Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions: Generating Knowledge Profiles from Resumes. *2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT)*, 97-106. <https://doi.org/10.1109/CONISOFT.2017.00019>.
- Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *ArXiv*, abs/2206.10498. <https://doi.org/10.48550/arXiv.2206.10498>.
- Veen, D., Uden, C., Blankemeier, L., Delbrouck, J., Aali, A., Blüthgen, C., Pareek, A., Polacin, M., Collins, W., Ahuja, N., Langlotz, C., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. (2023). Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3483777/v1>.
- Vijay, S., & Priyanshu, A. (2022). NERDA-Con: Extending NER models for Continual Learning - Integrating Distinct Tasks and Updating Distribution Shifts. *ArXiv*, abs/2206.14607. <https://doi.org/10.48550/arXiv.2206.14607>.
- Walford-Wright, G., & Scott-Jackson, W. (2018). Talent Rising; people analytics and technology driving talent acquisition strategy. *Strategic HR Review*. <https://doi.org/10.1108/SHR-08-2018-0071>.
- Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., & Kurohashi, S. (2023). GPT-RE: In-context Learning for Relation Extraction using Large Language Models. *ArXiv*, abs/2305.02105. <https://doi.org/10.48550/arXiv.2305.02105>.
- Wang, L., Near, J., Somani, N., Gao, P., Low, A., Dao, D., & Song, D. (2019). Data Capsule: A New Paradigm for Automatic Compliance with Data Privacy Regulations. , 3-23. https://doi.org/10.1007/978-3-030-33752-0_1.
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., ... & Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.
- Wang, W., Gu, Z., & Tang, K. (2020). An Improved Algorithm for Bert. *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*. <https://doi.org/10.1145/3444370.3444557>.
- Wang, X., Yang, C., & Guan, R. (2018). A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, 9, 373-382. <https://doi.org/10.1007/s13042-015-0426-6>.
- Wang, Y., Cui, L., & Zhang, Y. (2021). Improving Skip-Gram Embeddings Using BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1318-1328. <https://doi.org/10.1109/TASLP.2021.3065201>.

- Wiedemann, G., Yimam, S., & Biemann, C. (2020). UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection. , 1638-1644. <https://doi.org/10.18653/v1/2020.semeval-1.213>.
- Wild, S., Parlar, S., Hanne, T., & Dornberger, R. (2021, March). Naïve Bayes and Named Entity Recognition for Requirements Mining in Job Postings. In *2021 3rd International Conference on Natural Language Processing (ICNLP)* (pp. 155-161). IEEE.
- Xie, T., Li, Q., Zhang, Y., Liu, Z., & Wang, H. (2023). Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models. *ArXiv*, abs/2311.08921. <https://doi.org/10.48550/arXiv.2311.08921>.
- Xing, F., Cambria, E., & Welsch, R. (2017). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50, 49-73. <https://doi.org/10.1007/s10462-017-9588-9>.
- Xu, L., Xu, L., Jie, Z., Lu, W., & Bing, L. (2021). Better Feature Integration for Named Entity Recognition. *ArXiv*, abs/2104.05316. <https://doi.org/10.18653/V1/2021.NAAACL-MAIN.271>.
- Yang, C., Hou, Y., Song, Y., Zhang, T., Wen, J., & Zhao, W. (2022). Modeling Two-Way Selection Preference for Person-Job Fit. *Proceedings of the 16th ACM Conference on Recommender Systems*. <https://doi.org/10.1145/3523227.3546752>.
- Yang, H., Qin, Y., Deng, Y., & Wang, M. (2020). NMT Enhancement based on Knowledge Graph Mining with Pre-trained Language Model. *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, 185-189. <https://doi.org/10.23919/ICACT48636.2020.9061292>.
- Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ArXiv*, abs/2304.13712. <https://doi.org/10.48550/arXiv.2304.13712>.
- Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.
- Yao, Y., Dong, B., Zhang, A., Zhang, Z., Xie, R., Liu, Z., Lin, L., Sun, M., & Wang, J. (2022). Prompt Tuning for Discriminative Pre-trained Language Models. , 3468-3473. <https://doi.org/10.48550/arXiv.2205.11166>.
- Young, J. (2013). Middle-skill jobs remain more common among rural workers. . <https://doi.org/10.34051/P/2020.196>.
- Yu, G., Yang, Y., Wang, X., Zhen, H., He, G., Li, Z., Zhao, Y., Shu, Q., & Shu, L. (2020). Adversarial active learning for the identification of medical concepts and

- annotation inconsistency. *Journal of biomedical informatics*, 103481 .
<https://doi.org/10.1016/j.jbi.2020.103481>.
- Yu, J., Ji, B., Li, S., Ma, J., Liu, H., & Xu, H. (2022). S-NER: A Concise and Efficient Span-Based Model for Named Entity Recognition. *Sensors (Basel, Switzerland)*, 22. <https://doi.org/10.3390/s22082852>.
- Yu, S., Su, J., & Luo, D. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. *IEEE Access*, 7, 176600-176612. <https://doi.org/10.1109/ACCESS.2019.2953990>.
- Zafarian, A., Rokni, A., Khadivi, S., & Ghasifard, S. (2015). Semi-supervised learning for named entity recognition using weakly labeled training data. *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 129-135. <https://doi.org/10.1109/AISP.2015.7123504>.
- Zhang, H., Liu, X., & Zhang, J. (2023). Extractive Summarization via ChatGPT for Faithful Summary Generation. , 3270-3278. <https://doi.org/10.48550/arXiv.2304.04193>.
- Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. *International journal of medical informatics*, 132, 103985 . <https://doi.org/10.1016/j.ijmedinf.2019.103985>.
- Zhao, M., Javed, F., Jacob, F., & McNair, M. (2015). SKILL: A System for Skill Identification and Normalization. , 4012-4018. <https://doi.org/10.1609/aaai.v29i2.19064>.
- Zhou, W., Zhang, S., Gu, Y., Chen, M., & Poon, H. (2023). UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *ArXiv*, abs/2308.03279. <https://doi.org/10.48550/arXiv.2308.03279>.
- Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2019). Evaluating Commonsense in Pre-trained Language Models. , 9733-9740. <https://doi.org/10.1609/AAAI.V34I05.6523>.
- Zhuo, T., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. *ArXiv*, abs/2301.12867. <https://doi.org/10.48550/arXiv.2301.12867>.