UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOCRITOS"
MSC PROGRAMME IN DATA SCIENCE

Text analytics approaches to multichannel information summarisation on Fintech customers

by

Zoi Papakonstantinou

A thesis submitted in partial fulfillment
of the requirements for the MSc
in Data Science

Supervisor: Elias Zavitsanos

Co-supervisors: Konstantinos Bougiatiotis

Athens, December 2024

Text analytics approaches to multichannel information summarisation on Fintech customers

Zoi Papakonstantinou

MSc. Thesis, MSc. Programme in Data Science

University of the Peloponnese & NCSR “Democritos”, December 2024

UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOCRITOS"
MSC PROGRAMME IN DATA SCIENCE

Text analytics approaches to multichannel information summarisation on Fintech customers

by

Zoi Papakonstantinou

A thesis submitted in partial fulfillment
of the requirements for the MSc
in Data Science

Supervisor: Elias Zavitsanos

Co-supervisors: Konstantinos Bougiatiotis

Approved by the examination committee on December, 2024.

(Signature)	(Signature)	(Signature)
George Giannakopoulos	Christos Tryfonopoulos	Elias Zavitsanos

Athens, December 2024

UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOCRITOS"
MSC PROGRAMME IN DATA SCIENCE

Declaration of Authorship

(1) I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

(2) I confirm that this thesis presented for the degree of Master of Science in Informatics and Telecommunications, has

(i) been composed entirely by myself

(ii) been solely the result of my own work

(iii) not been submitted for any other degree or professional qualification

(3) I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification except as specified.

(Signature)

.....

Zoi Papakonstantinou

Athens, December 2024

Acknowledgments

I would like to express my deepest gratitude to those who have been instrumental in the completion of this thesis.

First, I am very grateful to my supervisor, Dr. Elias Zavitsanos, for his guidance, support, and encouragement throughout my research. His advice and feedback have been very helpful in improving this work.

I would also like to thank Konstantinos Bougiatiotis, whose help and support were essential during the preparation of this thesis. His help and dedication have been greatly appreciated.

My thanks extend to Qualco SA for recommending this topic and supporting my studies and research.

A special thanks to my colleagues and friends for their support and encouragement during this journey. Their friendship and understanding have been a great source of motivation.

Finally, I want to thank my family for their ongoing support and patience. Their belief in me and encouragement have been key in completing this work.

Thank you all for your invaluable contributions and support.

To my family.

Abstract

This thesis focuses on the design and implementation of a system that applies machine learning techniques to financial documents with the aim of automating their summarization. The research primarily utilized the K-Means algorithm to extract key sentences for the summary of each text document in the dataset provided by Qualco SA. Additionally, the proposed method was then applied to the Financial Narrative Summarization (FNS) 2023 dataset, where it demonstrated promising results in summarizing financial narratives. The models were evaluated using metrics such as ROUGE scores to assess their effectiveness in capturing key information from the documents.

Περίληψη

Η παρούσα εργασία επικεντρώνεται στον σχεδιασμό και την υλοποίηση ενός συστήματος που εφαρμόζει τεχνικές μηχανικής μάθησης σε χρηματοοικονομικά έγγραφα με στόχο την αυτόματη περίληψη αυτών. Στο πλαίσιο της έρευνας, χρησιμοποιήθηκαν διάφορα μοντέλα, με κυριότερο τον αλγόριθμο K-Means, ο οποίος χρησιμοποιήθηκε για την επιλογή προτάσεων από το σύνολο δεδομένων της Qualco SA για την παραγωγή περιλήψεων. Η προτεινόμενη μέθοδος εφαρμόστηκε στη συνέχεια στο dataset του Financial Narrative Summarization (FNS) 2023, επιδεικνύοντας υποσχόμενα αποτελέσματα στη σύνοψη χρηματοοικονομικών αφηγήσεων. Τα μοντέλα αξιολογήθηκαν με μετρικές όπως η βαθμολόγηση ROUGE για την εκτίμηση της αποτελεσματικότητάς τους στην αποτύπωση των βασικών πληροφοριών από τα έγγραφα.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the contemporary digital era, individuals are bombarded with a constant influx of information across various channels, from social media to financial news. With the sheer volume of data generated daily, it becomes imperative to develop effective methods to sift through this information and extract meaningful insights. Text analytics, particularly text summarization, has emerged as a critical tool in this process, enabling the condensation of vast amounts of text into digestible summaries that retain essential information.

NLP emerged during the Second World War (1940s) as "Machine Translation (MT)" to translate the Russian language into English and vice versa with the help of a computer. Although there have been developments in the language of syntactic theory and parsing algorithms over the following decades none have been sufficient to create an efficient MT. As research in the field of technical summarization has progressed, many developments have been made in understanding or interpreting the meaning of sentences in a body of textual data. An important development in the field of natural language processing has been text summarization using neural networks [1].

Two principal approaches have emerged in text summarization: extractive and abstractive summarization. Extractive summarization involves selecting key sentences or phrases from the original text to create a summary. In contrast, abstractive summarization aims to generate summaries that are more coherent and closer

to human-like understanding by paraphrasing and synthesizing the content. While extractive methods have shown effectiveness in specific domains, such as news articles and technical documents, abstractive summarization presents a more advanced challenge, seeking to produce summaries with higher relevance and reduced repetition.

The motivation for this research stems from the growing demand for efficient summarization techniques in the Fintech sector. Financial technology (Fintech) companies deal with diverse data sources, including customer feedback, transaction records, and market analysis reports. The ability to summarize this multifaceted information effectively can significantly enhance decision-making processes, customer experience, and strategic planning. The increasing complexity and volume of data in Fintech underscore the need for robust summarization methods that can handle various data types and sources.

This thesis explores text analytics approaches to multichannel information summarization, specifically focusing on Fintech customers. The study leverages data provided by Qualco, applying three distinct models: K-means clustering, TF-IDF (Term Frequency-Inverse Document Frequency), and a random sentences approach. Additionally, the K-means model is further evaluated on data from the FNS 2023 dataset. The contributions of this research are as follows:

- **Subset of Gold Summaries for Evaluation:** A subset of gold standard summaries was created from the Qualco dataset. This subset served as a critical benchmark for evaluating and comparing the performance of the proposed summarization techniques, providing a reliable reference point for robust and meaningful evaluation.
- **Model Evaluation:** Assessment of the performance of K-means clustering, TF-IDF, and random sentence approaches in summarizing Fintech-related data, providing insights into their effectiveness in different contexts.
- **Application of K-means:** Application and evaluation of the K-means clustering model on a specialized dataset (FNS 2023), offering a comparative analysis of its performance in summarizing Fintech data.

- **Practical Insights:** Generation of practical recommendations for Fintech companies on leveraging text summarization techniques to enhance their data processing capabilities and improve customer engagement.

By addressing these areas, this research aims to contribute valuable knowledge to the field of text analytics and summarization, with a particular focus on the Fintech sector's unique needs and challenges.

1.1 Problem description

The financial technology sector (Fintech) is increasingly dependent on sophisticated data analysis to better understand and manage customer relationships, particularly in areas such as debt management services. One of the primary challenges is dealing with the large amounts of unstructured data generated through various customer interactions. This data comes from sources such as digital documents, emails, text messages, and call center logs, which contain valuable insights but are difficult to process and summarize effectively.

In this context, the problem is twofold: first, how to efficiently summarize these unstructured data to extract actionable insights; and second, how to ensure that the summaries provide a nuanced and comprehensive understanding of customer behavior and history. Fintech companies rely on these insights to make critical decisions that can significantly impact customer relations and service delivery.

To address these challenges, this thesis utilizes multiple datasets, including the newly introduced FNS2023 dataset, which offers additional context and depth to the analysis. The integration of this dataset is crucial for developing more accurate and informative summaries that capture the full spectrum of customer interactions.

1.2 Thesis structure

This thesis is organized into seven chapters, each addressing different aspects of the research:

- **Chapter 2: Related Work** - Provides an in-depth review of existing literature on text summarization, with a particular focus on methods applicable to

the Fintech sector. This chapter also explores the specific challenges associated with summarizing data from various sources.

- **Chapter 3: Data** - Discusses the datasets utilized in this research, including the primary data sources and the FNS2023 dataset. This chapter details the pre-processing steps taken to prepare the data for analysis and the unique characteristics of each dataset.
- **Chapter 4: Method/Methodology** - Outlines the approaches and algorithms employed in this study, including K-Means clustering, TF-IDF, and random sentence selection techniques for creating summaries. The chapter explains how these methods were adapted to handle the complexities of datasets.
- **Chapter 5: Experimental Setup** - Describes the experimental design, including the procedures followed to evaluate the performance of different summarization techniques. This chapter also explains how the inclusion of the FNS2023 dataset enhances the robustness of the experimental results.
- **Chapter 6: Results/Discussion** - Presents and discusses the results of the experiments, comparing the effectiveness of various summarization methods. The chapter also examines how the integration of the FNS2023 dataset influenced the outcomes and provided deeper insights.
- **Chapter 7: Conclusions** - Summarizes the key findings of the research, discusses their implications for the Fintech industry, and suggests areas for future research. The conclusions highlight the importance of using diverse datasets, like FNS2023, for comprehensive customer analysis.

1.3 Thesis Topic summary

This thesis aims to use various sources of free text information, including digital documents, emails, text messages, and calls to call center agent comments, to generate comprehensive, summarized 360-degree views of customer history and behavior within the Fintech industry. Specifically, it focuses on customers involved in debt services, where understanding the full scope of customer interactions is crucial for effective decision-making.

By integrating and analyzing data from multiple communication channels, including the newly introduced FNS2023 dataset, this research develops methods to summarize and consolidate this information into cohesive insights. These insights are intended to provide Fintech companies with a holistic understanding of their customers, allowing for more personalized service delivery and improved customer relationship management.

The research highlights the importance of utilizing diverse and rich datasets to create accurate and nuanced summaries that capture the complexity of customer behaviors. In doing so, it contributes to the broader field of text analytics, offering valuable tools and methodologies for the Fintech sector to better engage with and understand their customers.

Chapter 2

Related Work

2.1 Brief History of Text Summarization

Text summarization is an important application of Natural Language Processing (NLP). NLP emerged during the Second World War (1940s) as "Machine Translation (MT)" to translate the Russian language into English and vice versa with the help of a computer. Although there have been developments in the language of syntactic theory and parsing algorithms over the following decades, none have been sufficient to create an efficient MT. However, significant progress has been made since then in modeling sentence meaning, leading to advancements in text summarization, especially using neural networks [1].

Text summarization has remained a subject of interest in subsequent decades, with Hans Peter Luhn [2] introducing the concept of automatic text summarization in 1958. Luhn, a computer scientist, became a pioneer in this field, focusing on creating concise summaries by identifying and extracting key phrases, termed "important words". His innovative approach involved ranking words based on their frequency in the text and selecting the most significant ones for summarization. Luhn's work laid the groundwork for subsequent research in text summarization and information retrieval. Over time, the field has advanced, integrating diverse techniques and methodologies to enhance the accuracy and efficiency of automatic text summarization systems. In the late 1960s, Harold P. Edmundson [3] conducted a major

research project that used methods based on the presence of specific author words, title words appearing in the text, and sentence positions to extract meaningful sentences for the text summary. Since then, numerous influential and exciting studies have been conducted to address the challenge of automatic text summarization.

The evolution of text summarization has seen advancements from extractive unsupervised statistical approaches to abstractive supervised methods utilizing deep learning models (Liu, 2019 [4]). Its impact on Natural Language Processing (NLP) is profound, particularly in today's digital age, where time constraints make it impractical to thoroughly read lengthy documents. Text summarization becomes essential in condensing voluminous data, aiding in efficient information tracking across various fields.

2.1.1 Text Summarization Methods

We can categorize text summarization into two primary approaches: those based on extraction and those on abstraction. The **extractive text summarization** approach involves extracting essential words from an original document and combining them to create a summary. In practice, it involves the selection of particular sentences to be included in the final summary. Extractive summarization uses a scoring mechanism to rank the relevance of phrases to select just those that are most relevant to the source document's meaning. This technique extracts the required text according to the specified criteria without making significant changes to the documents. The extractive summarization method works with the help of algorithms such as LexRank, Luhn and LSA, among other.

On the other hand, **abstractive summarization** focuses on the most critical information in the original text and creates a new set of sentences for the summary. The new sentence might not be a part of the source text. This approach differs substantially from the extractive text summarization approach, as the latter generates a summary based on selecting sentences from the original text. The abstraction summarization technique entails identifying key pieces, interpreting the context, and re-creating them in a new way. It aims to capture the most crucial information in

the shortest possible text. The abstractive summarization method works well with deep learning models like sequence-to-sequence (e.g: LSTM) [5].

Multi-document summarization involves condensing information from multiple sources into a concise summary, capturing key points from various documents on the same topic. On the other hand, single-document summarization focuses on condensing information from a single document into a shorter version while retaining the essential content.

In the context of multilingual summarization, the goal is to generate summaries in multiple languages, allowing users to access information in their preferred language. This involves challenges such as dealing with language-specific nuances and ensuring the cross-lingual applicability of summarization methods. In contrast, mono-lingual summarization focuses on generating summaries within a single language, without the added complexity of translation and language-specific issues. Both multi-document and multilingual summarization present unique challenges and opportunities for researchers in the field of natural language processing and information retrieval, aiming to provide users with comprehensive and accessible summaries across different languages and sources [6].

In an overview, text summarization involves creating concise summaries while retaining key information and content value. The language of the summary should be unambiguous, conveying meaning to the reader. The demand for machine learning algorithms that can quickly and accurately summarise long texts is high. In this work, we focus on extractive summarization, and therefore we provide an overview of the domain in the following sections.

2.2 Extractive Summarization

As already mentioned, extractive summarization is a text summarization technique focused on condensing content by selecting important sentences or passages from the original document [7][8][9][10]. The objective is to create a shorter version that retains the essential meaning and structure while capturing the key information and main points. Several key points of extractive summarization are:

- **Selection of important content**

  Extractive summarization involves identifying and extracting sentences or passages that are deemed important based on various criteria such as relevance, importance, and informativeness.

- **Use of statistical and linguistic features**

  The importance of sentences in extractive summarization is determined using statistical and linguistic features of the text. Features like content words (keywords), sentence location, and sentence length may influence the selection process.

- **Methods and techniques**

  Extractive summarization methods may utilize the Term Frequency-Inverse Document Frequency (TF-IDF) to score sentences based on their relevance to the overall document. Sentence weighting and selection algorithms are used to determine which sentences should be included in the final summary. It also addresses challenges encountered in extractive summarization, such as maintaining coherence and avoiding redundancy in the summary.

- **Process**

  The extractive summarization process typically involves pre-processing steps like identifying sentence boundaries and eliminating stop words, followed by processing steps where sentences are scored and selected for inclusion in the summary.

- **Challenges**

  Extractive summarization faces challenges such as maintaining coherence, avoiding redundancy, and accurately capturing essential information from the source text. Issues like including conflicting information, handling pronouns and temporal expressions, and ensuring overall coherence can impact the quality of the extractive summary.

In extractive summarization, as well as the important categories mentioned above, statistical and linguistic factors play a crucial role in determining the meaning of sentences. Among the main factors that influence the summary process are

the following:

- **Content words (or keywords)**
  Statistical analysis of content words, typically nouns, using measures like term frequency-inverse document frequency (TF-IDF), helps identify important keywords in the text. Sentences containing these keywords are more likely to be included in the summary.

- **Sentence location**
  The position of a sentence within the document can indicate its importance. For example, sentences appearing in the introduction or conclusion may carry more weight and relevance than those in the middle of the text.

- **Sentence length**
  Statistical analysis of sentence length can be a crucial factor in determining importance. Longer sentences may contain more information, but shorter sentences may be more concise and to the point, influencing their selection for the summary.

- **Discourse analysis**
  Linguistic features related to discourse analysis, such as the flow of the author's argument and the overall discourse structure of the text, can help in identifying sentences that are central to the main message. Removing peripheral sentences can improve the coherence and relevance of the summary.

- **Occurrence of redundant information**
  Linguistic markers like speech markers ("because", "furthermore") at the beginning of sentences can indicate non-essential information. Algorithms can use this linguistic feature to filter out sentences that may not contribute significantly to the summary.

By considering these statistical and linguistic factors, extractive summarization algorithms aim to effectively evaluate the importance of sentences in a document and select the most relevant ones for inclusion in the summary.

2.2.1 Technical Approaches

In the broader field of extractive summarization, several research efforts have been proposed, combining various methodologies and techniques. Each study offers a unique perspective, presenting a variety of approaches to the task. From traditional statistical models to innovative machine learning algorithms, researchers are constantly pushing the boundaries of text summarization.

Before delving into a detailed examination of the techniques employed in text summarization, it is crucial to understand the various approaches adopted in this field. Each technique to be analyzed serves as a tool aimed at effectively extracting essential information from text, employing different methods and algorithms. Through in-depth analysis, we reveal the strategies used to deal with the complexities of text summarization. This section will explore the main techniques proposed in the literature, examining how each contributes to natural language processing and information retrieval, ultimately aiming to optimize text summarization.

1. **Statistical Methods** [11][12]refer to techniques that utilize statistical data to identify and rank the most significant sentences or words within a text. These methods often include:

   - **Frequency Methods**

     These techniques are based on the frequency of word appearances in the text. The primary concept is that words appearing more frequently are likely to be more significant. A simple application of this method is to count the frequency of each word and select sentences that contain the most frequent words for the summary.

   - **TF-IDF (Term Frequency-Inverse Document Frequency)**

     TF-IDF [13] is a development over simple frequency methods, offering a more complex and refined approach. While it still relies on the frequency of words, TF-IDF goes a step further by also calculating the rarity of the word across a set of documents or corpus. This measure reflects how important a word is in a document not only based on its frequency but also its scarcity across a broader set of documents. This helps to highlight

words that are uniquely significant for a particular document, reducing the influence of commonly frequent but less informative words.

2. **Linguistic Methods** [14] involve the analysis of linguistic elements of the text, focusing on structure, syntax, and semantics to identify the most significant sentences.

   - **Thematic Cohesion**

     Thematic Cohesion focuses on how sentences and paragraphs are organically integrated into the text to create a coherent and consistent narrative. This approach seeks to follow the flow of ideas and the progressive development of themes within the text, identifying sentences that manage to maintain the core of the discussion and advance the understanding of the central theme.

   - **Recognition of Nominal Phrases and Entities**

     This method involves extracting and assessing the importance of nominal phrases and entities, such as names of individuals, organizations, geographic locations, etc. In extractive summarization, sentences that include such information are often considered crucial as they contain specific and significant data that contribute to understanding the main idea or events presented in the text.

3. **Graph-based Methods** [9] are a technique for extractive text summarization that involves representing text as a graph, where sentences are nodes and their relationships are edges.

   - **PageRank**

     This algorithm was originally developed to evaluate the importance of webpages on the Internet. In text summarization, PageRank is used to determine the significance of each sentence based on the number and quality of links entering and exiting it. Sentences that receive a high ranking are considered more important and are more likely to be included in the summary.

   - **Other Ranking Algorithms**

Besides PageRank, there are other algorithms that can be applied to graphs for ranking sentences, such as HITS (Hyperlink-Induced Topic Search), TextRank (a variation of PageRank adapted for texts), and LexRank. Each algorithm has different priorities and methods for evaluating the relationships between sentences, offering a variety of tools for selecting the best content for summarization.

This variety of algorithms allows researchers and developers to choose and tailor the most suitable method depending on the needs of the text and the specific requirements of their application.

In Fig. 2.1, the graph-based clustering method [15] for document analysis is depicted. Initially, features are extracted from documents, and a similarity graph is constructed with these features. This graph is then partitioned into clusters, each representing a group of closely connected documents. This method effectively organizes documents into distinct categories based on common characteristics. By integrating the PageRank algorithm into the analysis, we can further evaluate the importance of each document or cluster, based on the quality and number of connections among them. This allows for the ranking of documents according to their influence within the dataset, providing a deeper understanding of the relationships and the value of the information they contain.

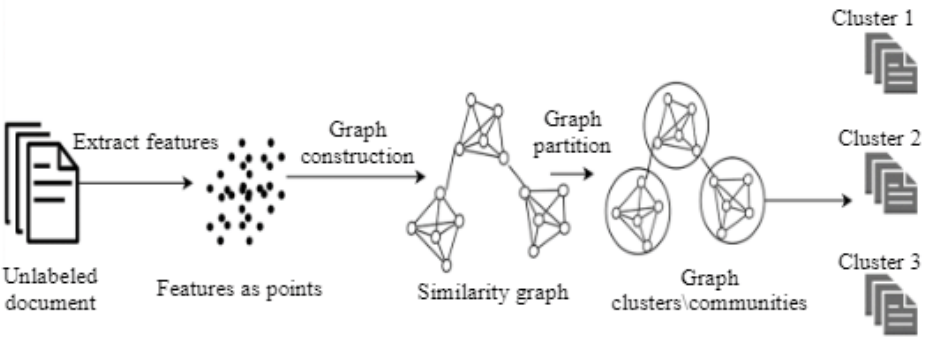


Figure 2.1: Graph-based clustering method.

4. **Machine Learning Methods** [16] refer to techniques that utilize mathematical models and algorithms to analyze texts and identify the most significant sentences or sections. These methods are primarily categorized into two types:

supervised learning, where models are trained with previously labeled data, and unsupervised learning, where algorithms seek to discover information and create groupings without predefined labels.

- **Supervised Learning**

  Supervised learning is used when we have a dataset with labeled information, which means it is already determined which sentences are important and which are not. This allows the model to "learn" by recognizing features that make a sentence significant. Models such as neural networks, logistic regression, or support vector machines (SVM) can be trained on this type of data to predict the significance of new sentences in unpublished texts.

- **Unsupervised Learning**

  Unsupervised learning involves algorithms that analyze data without requiring any labeled datasets. In extractive text summarization, this means the algorithms must infer the structure and importance of text segments based solely on the data's inherent features and patterns. This category of learning is broad and includes various techniques, each with unique ways of handling and interpreting data. One specific technique under unsupervised learning is clustering, which groups data points (in this case, sentences) that are similar to each other into clusters. Each cluster of sentences is expected to represent a unique theme or topic present in the text. Among the various clustering algorithms, k-means is particularly noteworthy.

  k-means clustering is a popular method that partitions n items into k clusters in which each item belongs to the cluster with the nearest mean. This method effectively organizes large volumes of text data because it efficiently categorizes sentences into distinct thematic groups based on their similarity. In the context of extractive text summarization, k-means facilitates the selection of representative sentences from each cluster to compile a comprehensive summary of the text. By leveraging k-means, summarization systems can handle diverse datasets without pre-tagged

data, making it an ideal choice for applications without prior annotations. In Fig. 2.2 illustrates the journey of unsupervised learning clustering using the k-means algorithm, from data collection and cleaning to the final grouping of documents [17] .

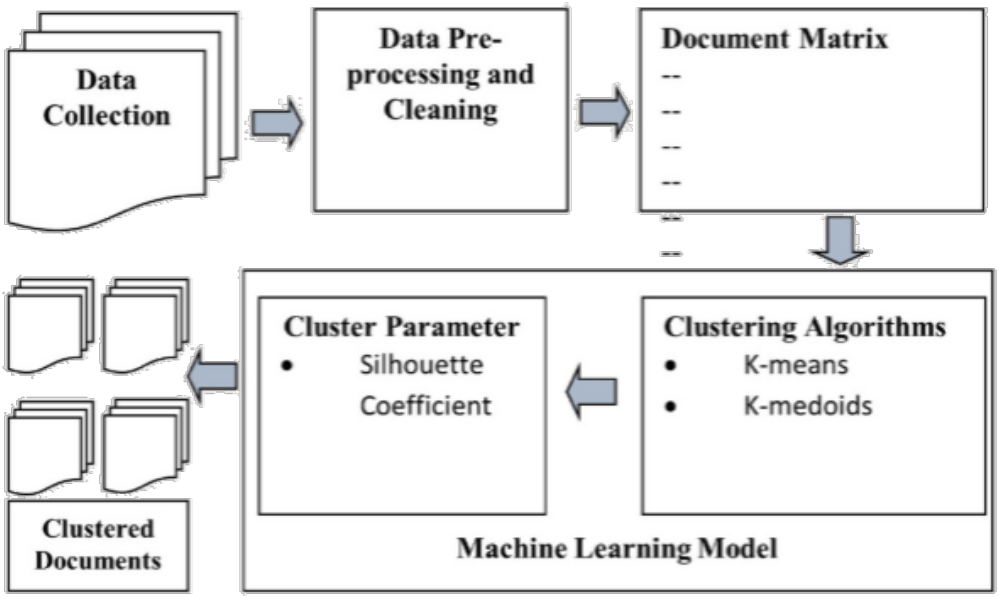


Figure 2.2: Unsupervised Learning (Clustering).

This approach not only streamlines the summarization process but also enhances the adaptability of systems to new, unstructured text inputs, underscoring the utility of k-means in unsupervised learning scenarios for natural language processing.

5. **Hybrid Methods** [4][18][19] to the combination of multiple techniques to capitalize on the strengths and minimize the weaknesses of individual approaches. This combination is particularly useful because:

   - **Increased Accuracy:** Using multiple techniques can improve the overall accuracy of the summarization, as weaknesses in one method can be offset by strengths in another.
   - **Coverage of More Themes:** Some methods might be better at extracting certain types of information (e.g., central ideas, key events) than others, which might be more effective in highlighting details or technical

data.

- **Flexibility:** Hybrid systems offer greater flexibility to adapt to various kinds of texts and different summarization requirements.
- **Error Resilience:** Combining techniques can also make the summarization system more resilient to errors. If one method fails or is not as effective, other methods can ensure that the overall quality of the summary is not significantly affected.

The choice and combination of appropriate methods depend on the specific needs of the application and the nature of the text being used. In a thesis, a hybrid approach can be presented as an innovative solution that combines technological innovation with accuracy and efficiency.

As previously discussed, extractive text summarization is an essential task in natural language processing (NLP) that aims to reduce large textual documents into succinct summaries, preserving crucial information. This process is particularly challenging and interesting when applied to the Greek language, due to its unique linguistic characteristics including complex morphology, syntax, and semantics.

In the study by Kantzola Evangelia [8], a thorough review is provided on extractive text summarization techniques specifically adapted for the Greek language. This research explores various methodologies and approaches, addressing the distinct challenges involved in summarizing Greek texts. It includes the development and evaluation of extractive summarization models, the use of machine learning and linguistic methods for sentence selection, and the examination of language-specific preprocessing techniques. Furthermore, the paper highlights the importance of developing evaluation datasets and benchmarks tailored to the Greek linguistic context. It also emphasizes the significance of coherence, redundancy reduction, and linguistic quality in summaries of Greek texts.

The application of such techniques is pivotal within the broader linguistic landscape and particularly in the context of Greek text summarization. Despite their potential, the exploration of summarization methods specifically tailored to the Greek language remains a relatively unexplored field, presenting significant opportunities for future research.

The aforementioned techniques and methodologies are frequently applied in the context of extractive document summarization. Such approaches are particularly relevant as the work presented herein relies on similar ideas and builds upon these foundations.

2.3 Evaluation

Building on the summarization techniques discussed previously, it is essential to rigorously evaluate the effectiveness and accuracy of the generated summaries. Text summary evaluation involves assessing the quality of the summary using various metrics. These include ROUGE [20] and BLEU [21] for overlap measurement, the human evaluation focused on readability and coherence, content-based metrics for semantic similarity, and computational resource estimates. Given the multifaceted nature of summary quality, a comprehensive assessment approach incorporating a mix of metrics and human judgment is necessary. This section will delve into each evaluation method associated with extract summarization, discussing its relevance, methodology, and applications in the context of text summarization.

Numerous studies [7][8][9][22][23] provide comprehensive insights into the evaluation methods employed for assessing the effectiveness and quality of extractive summarization techniques. These works underline the critical importance of employing both intrinsic and extrinsic evaluation approaches to guarantee the coherence, relevance, and informativeness of the summaries produced. Intrinsic evaluation methods incorporate human judgment, where evaluators measure the quality of a summary against pre-defined criteria, such as coherence, focus, and structural integrity. This process ensures that summaries are not only readable but also serve their intended communicative purpose effectively. Extrinsic evaluation, in contrast, assesses the utility of summaries by their performance in practical tasks, such as enhancing the efficacy of information retrieval systems, thereby reflecting their real-world applicability.

Additionally, these studies stress the value of objective evaluation metrics, such as Spearman, Pearson, and Kendall correlation coefficients, which are utilized to compare system-generated summaries with human-crafted references. These statis-

tical tools are indispensable for verifying that summaries meet the specific information requirements outlined in topic statements.

The referenced papers also discuss the utilization of evaluation corpora and standardized metrics like ROUGE, which are pivotal for assessing the quality of automatic summarization systems by comparing machine-generated summaries to those crafted by humans. Some authors propose innovative evaluation measures, like the LSA-based evaluation method, which quantifies the similarity of main topics between system summaries and reference documents, offering a nuanced approach to ranking summarization systems.

In sum, these scholarly contributions emphasize the indispensable role of thorough and varied evaluation techniques in summarization research. By employing rigorous evaluation methodologies and contrasting outcomes with gold-standard references, these works significantly advance our understanding and development of summarization technologies. They ensure that generated summaries are not only concise and informative but also coherently encapsulate the critical information from the original texts, thus supporting effective communication and decision-making processes.

In light of the robust methodologies discussed, this dissertation will employ the ROUGE metric to ensure that our evaluation of extractive summarization techniques is both precise and aligned with established standards, facilitating a reliable assessment of their effectiveness in handling Greek financial texts.

2.4 Financial Narrative Summarisation (FNS 2023)

In recent years, the automation of financial text analysis has gained unprecedented momentum, driven by advances in Natural Language Processing (NLP) and the growing complexity of financial data. The 5th Financial Narrative Processing Workshop (FNP 2023), held as part of the 14th Edition of the Language Resources and Evaluation Conference, has been at the forefront of this evolving field. This workshop provides a critical platform for researchers and practitioners to exchange knowledge, discuss challenges, and showcase the latest innovations in financial narrative summarization.

FNP 2023 addressed several key areas within the domain of financial narrative processing, emphasizing the importance of sophisticated text analysis techniques to manage, summarize, and interpret large volumes of financial documents. The workshop featured shared tasks that are pivotal in pushing the boundaries of what automated systems can understand and process. These tasks included:

- **Text Summarization:** Both extractive and abstractive techniques were explored, focusing on how to effectively condense financial narratives while retaining crucial information. This task is essential as it helps stakeholders quickly grasp the essence of financial documents without delving into detailed reports.
- **Structure Detection:** This involves identifying and categorizing the structural elements of financial texts, which is crucial for automated systems to effectively parse and interpret complex documents.
- **Causal Sentence Detection:** Detecting causality in texts is particularly challenging yet vital for drawing meaningful conclusions from financial reports and disclosures.

The exploration of different summarization techniques at the workshop sheds light on the ongoing efforts to enhance how financial information is processed and presented. With the financial sector increasingly relying on timely and accurate data summarization, these advancements are critical in supporting decision-making processes within businesses and regulatory bodies.

The primary impetus for this research stemmed from the increasing complexity and volume of financial narratives in Greek language reports, which present unique challenges in automatic text summarization. Greek financial documents often exhibit less structured formats compared to their English counterparts, complicating the extraction and summarization processes. This research [24] aims to address these challenges by leveraging advanced Natural Language Processing (NLP) techniques to enhance summarization efficiency and accuracy, thereby assisting stakeholders in making informed financial decisions without navigating through voluminous reports.

Key techniques employed include a combination of extractive and abstractive summarization methods. Extractive techniques, such as TF-IDF scoring combined

with clustering algorithms, were used to identify and rank the most pertinent information within the texts. Simultaneously, abstractive methods involved the use of fine-tuned multilingual models like mT5 and T5, which can generate concise summaries that capture the essence of financial narratives without directly copying the source texts. Additionally, the application of translation-summary-translation strategies for Greek reports ensures that the nuances of financial language are retained, catering to the specific challenges posed by less structured data formats. These sophisticated approaches not only streamline the summarization process but also improve the quality of the generated summaries, making them more useful for quick decision-making.

These sophisticated approaches not only streamline the summarization process but also improve the quality of the generated summaries, making them more useful for quick decision-making. To assess the effectiveness of these summarization techniques, the research utilized the ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation), which measures the overlap of n-grams between the automated summaries and expert-written gold standards. This evaluation method helps to ensure that the summaries are not only succinct but also accurate and reflective of the content's critical points. Furthermore, feedback from stakeholders in the financial sector indicates that these automated summaries significantly enhance their ability to quickly grasp essential financial insights without delving into lengthy documents. This benefit is crucial for timely and informed decision-making in fast-paced financial environments.

This research aligns closely with my work, as it delves into financial narrative summarization and supports the Greek language, two fundamental aspects of my academic and professional focus. The utilization of advanced NLP techniques in Greek financial documents demonstrates a direct correlation with my research interests, making it a pivotal reference for my studies. However, my approach diverges in critical ways due to unique challenges that require distinct strategies. For instance, my work places a significant emphasis on the temporal dimension of financial data. Recognizing the importance of timeliness in financial reporting, my methodology prioritizes recent information, selectively incorporating the most current data into

summaries to ensure they reflect the latest financial landscape. This aspect is particularly crucial as it enhances the decision-making process for stakeholders who rely on the most up-to-date information to make informed financial decisions. Thus, while the foundational techniques of this research provide a strong basis for understanding and developing NLP applications in finance, my work extends these by addressing specific, time-sensitive needs in financial summarization.

Chapter 3

Data

This chapter focuses on the QUALCO dataset, which includes digital documents, emails, text messages, and call center agent comments from a fintech company specializing in debt services. This data provides a comprehensive view of customer history and behavior, serving as the primary source for developing our automated text summarization models. Additionally, we used data from the Financial Narrative Summarisation Shared Task (FNS 2023), which includes UK, Greek, and Spanish annual reports. This dataset, part of the 5th Financial Narrative Processing Workshop (FNP 2023), helped us evaluate the performance of our models on diverse financial documents. Combining these datasets allowed for a thorough assessment of our summarization techniques.

3.1 Qualco SA Dataset

This thesis integrates data sourced from Qualco S.A. (Qualco). The data is obtained through the bank's call center where customers manage their debt-related process. Qualco provided a dataset that includes information on three entities, namely *Customer*, *Account*, and *Actions*. The thesis focuses specifically on the Actions, which include customer details, deposit numbers, update dates, and comments recorded by staff during customer interactions. An example illustration of 10 such records is shown in Fig. 3.1.

```
# Display data
Actions_Data.head(10)
```

	Action_ID	Snapshot_Number	Action_Date	Debt_ID	Customer_ID	Action_Comment
0	151924128	38.0	2019-02-01	510056.0	262177.0	3/4/##########
1	151924066	38.0	2019-02-01	439243.0	312593.0	10/4/##########
2	151926743	38.0	2019-02-01	362644.0	234812.0	ειρ. ιλιου 2/6/#### δεκτη εν [xxxx] συνολικη απαιτηση [xxxx]: -### #######,
3	151926742	38.0	2019-02-01	649999.0	234812.0	ειρ. ιλιου 2/6/#### δεκτη εν [xxxx] συνολικη απαιτηση [xxxx]: -### #######,
4	151926754	38.0	2019-02-01	649999.0	234812.0	ειρ. ιλιου 9/6/#### συνολικη απαιτηση [xxxx]: -### #######,
5	151926755	38.0	2019-02-01	362644.0	234812.0	ειρ. ιλιου 9/6/#### συνολικη απαιτηση [xxxx]: -### #######,
6	151926739	38.0	2019-02-01	686660.0	357297.0	κτχ θα κτθ 04/02
7	151926719	38.0	2019-02-01	369508.0	352021.0	### φραγη // ### κα [xxxx] [xxxx] απων ειναι εκτοσ [xxxx]
8	151926718	38.0	2019-02-01	639384.0	352021.0	### φραγη // ### κα [xxxx] [xxxx] απων ειναι εκτοσ [xxxx]
9	151926717	38.0	2019-02-01	497868.0	352021.0	### φραγη // ### κα [xxxx] [xxxx] απων ειναι εκτοσ [xxxx]

Figure 3.1: Examples showcasing 10 records from the Actions dataset.

The data sample consists of approximately **8 million records** belonging to **186,942 unique customers**, covering different periods and types of financial transactions. Furthermore, there are **311,036 debts**, indicating that on average each customer has approximately two debts. Specifically, the available details in each record are the following:

- **Action_ID:** Each record has a unique action identifier.
- **Snapshot_Number:** Each record is defined by a snapshot (datetime) number, denoting the month the data was provided.
- **Action_Date:** Date of creation of the comment.
- **Debt_ID:** Unique ID of the corresponding Debt that this action refers to.
- **Customer_ID:** Unique ID of the Customer to which the Debt belongs.
- **Action_Comment:** A comment describing the communication between a Qualco agent and the customer, **in Greek**.

To understand how our customer base is distributed, we investigated the distribution of customers based on the number of debts they have.

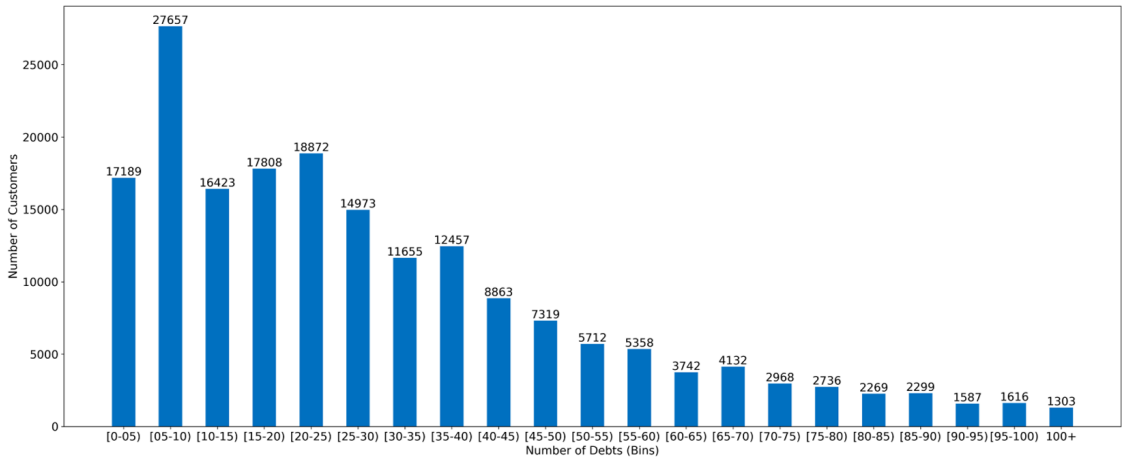


Figure 3.2: Frequency of debts per customer using a binning every 5 comments.

In Fig. 3.2 we observe that most customers fall within the second bin, indicating that 27,657 customers have between five and ten debts. Moreover, we visualize the number of words per comment in Fig. 3.3. It can be seen that the majority of comments consist of only one word. This word indicates that the discussion with the customer was not possible or no new comments were generated. Interestingly, one comment stands out with an extensive length of 256 words, showcasing a unique and detailed discussion compared to the rest of the dataset.

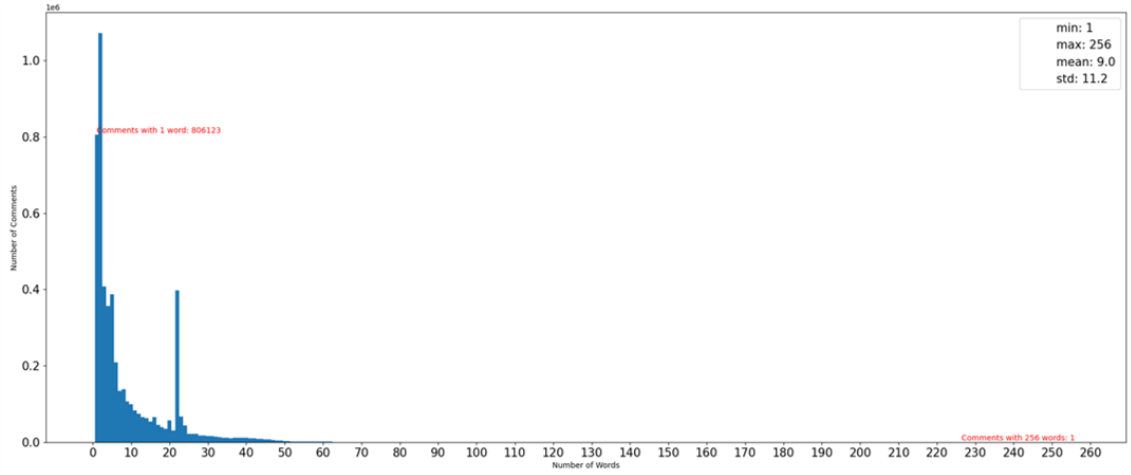


Figure 3.3: Frequency of words per comment.

The above observation highlights the variability in comment lengths and underlines the importance of this extreme situation for providing comprehensive feedback.

3.1.1 Pre-processing

Text pre-processing is a pivotal initial phase in natural language processing

(NLP) and text summarization tasks. This essential process revolves around converting raw textual data into a refined format, optimizing it for subsequent in-depth analysis and modeling. Employing various techniques, we enhance the quality and utility of the text data, that are riddled with noise as seen in the last column of Fig. 3.1. Some key pre-processing steps include tokenization, lowercasing, punctuation removal, stopword elimination, and stemming or lemmatization. By executing these steps, we achieve a more organized and coherent textual representation and lay the groundwork for extracting valuable insights and patterns through sophisticated analytical methods. We applied the following pre-processing steps to our data:

1. **Replace "[xxxx]" to null values**

   Removes redacted text placeholders for privacy compliance, such as under GDPR, preparing the data for accurate analysis.

2. **Remove punctuation characters**

   Punctuation is usually non-essential for many text-processing tasks like text classification or clustering. Removing it helps in standardizing the text, reducing complexity for algorithms.

3. **Convert to lowercase**

   This step ensures uniformity in the data by treating words like "Apple", "apple", and "APPLE" as the same word, which is crucial for most text analysis tools to function correctly.

4. **Remove digits**

   Unless numbers convey significant meaning relevant to the analysis, removing them helps focus on textual content, simplifying the data processing tasks.

5. **Remove extra spaces**

   Extra spaces can arise from errors in data entry or processing and do not contribute to meaningful data interpretation. Removing these helps in cleaning the dataset and ensures algorithms do not misinterpret separate terms.

6. **Remove stop words**

   Stop words such as "and", "the", and "is" often appear frequently in the text but offer little value in understanding the essence of the content. Their removal

increases the relevance of significant words in text analytics.

7. **Replace empty strings with NaN**

   Empty strings are treated as missing values and replacing them with NaN (Not a Number) standardizes these entries across the dataset, simplifying subsequent operations like data cleaning or filtering.

8. **Remove rows with NaN values**

   Removing rows containing NaN values helps in maintaining a dataset with complete information, which is essential for robust statistical analysis and machine learning modeling.

We noticed that our data set contained Greek words like $\delta\epsilon\nu$, which are important for our summary. Thus, we created a file with stop words and excluded the words $\delta\epsilon\nu$ or $\delta\epsilon$, in order not to change the meaning of the summary. Additionally, for GDPR compliance reasons, Qualco has implemented masking for clients' personal data, including sensitive information such as names, mobile numbers, email addresses, debt amounts, and more. This measure ensures that confidential information remains protected and secure, aligning with the strict data privacy regulations mandated by the GDPR. Subsequently, we provide a table with the data before and after pre-processing in Table 3.1.

Data Preprocessing Step	Before	After
Remove Punctuation Characters	κτχσ προτιθεται να αποπληρωσει εφαπαξ 2, *** € τα οποια θα συγκεντρωσει απο [χχχχ] προσωπα και θα τα εχει [χχχχ] αμεσα.	κτχσ προτιθεται να αποπληρωσει εφαπαξ 2 τα οποια θα συγκεντρωσει απο προσωπα και θα τα εχει αμεσα
Remove Numerals	κτχσ προτιθεται να αποπληρωσει εφαπαξ 2 τα οποια θα συγκεντρωσει απο προσωπα και θα τα εχει αμεσα	κτχσ προτιθεται να αποπληρωσει εφαπαξ τα οποια θα συγκεντρωσει απο προσωπα και θα τα εχει αμεσα

Table 3.1: Example application of the pre-procesing procedure.

Effective summarization hinges critically on the quality of text pre-processing. In the domain of financial narratives, where precision and clarity are paramount, removing extraneous elements such as punctuation and numbers not only makes the text more concise but also enhances readability, thereby facilitating a clearer

understanding of complex financial information. For instance, extracting keywords and significant phrases helps in distilling the essence of financial reports, enabling summarization models to focus more acutely on the core messages that are most relevant to stakeholders.

Moreover, implementing pre-processing steps such as stopword removal and case normalization brings uniformity and consistency to the text. These processes are essential for maintaining the analytical integrity of the text, ensuring that the summarization algorithms can interpret and analyze the content without the noise of irrelevant data. By standardizing the text input, pre-processing steps help in reducing the variability of the data, which in turn increases the accuracy and efficiency of the summarization outputs.

In summary, pre-processing creates a conducive environment that significantly enhances the performance of summarization techniques. This is especially critical in financial summarization where the stakes are high and the demand for accuracy and timely information is paramount. The pre-processed text allows for a more effective application of both extractive and abstractive summarization methods, ensuring that the summaries generated are not only informative and coherent but also reflective of the most pertinent financial data.

3.1.2 Details

In this section, we focus on visual representations to highlight patterns, trends, and points of interest within the data we analyze. Through graphs, charts, and other visualization methods, we aim to illuminate the key features of our dataset and provide deeper insights into the subject of our study.

Delving into the comments generated for each customer, we introduced a new column to calculate the number of comments per customer. Given the extensive volume of data, we aggregated the comments by both customer and debt, organizing them into 21 bins. These ranged from zero to the maximum number of comments recorded in the dataset, with each bin up to 100 further subdivided into increments of five for more granular analysis.

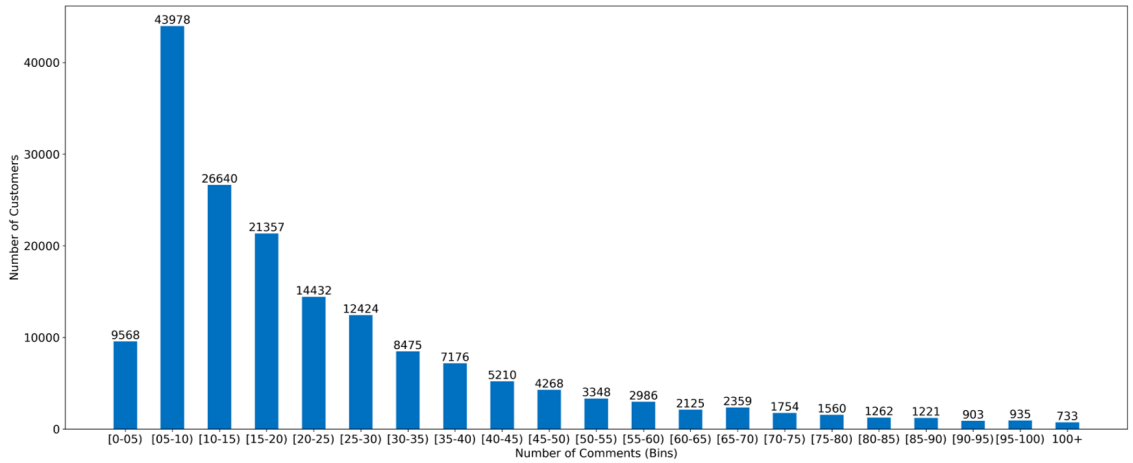


Figure 3.4: Frequency of comments per customer using a binning every 5 comments.

In Fig. 3.4, we observe that most customers are concentrated in the first few bins. Specifically, 43,978 customers have between five and ten comments. This distribution is anticipated, given the nature of customer comments during phone calls, which typically prioritize concise and direct communication. The concentration in lower bins suggests that most customer interactions are brief, potentially indicating efficient service processes or limited issues per call.

To further deepen our analysis, we also aggregated all comments per customer into a new metric that counts the number of sentences per customer. This measure offers insights into the length and complexity of interactions, providing a clearer picture of customer engagement levels and interaction styles. By analyzing sentence counts, we can identify patterns of communication that may suggest areas for improving customer service efficiency or targeting more detailed follow-up interactions where necessary.

As depicted in Fig. 3.5, we see two distinct visualizations that elucidate the distribution of customers across various sentence bins. The left side of the figure presents a pie chart, which colorfully illustrates the percentage of customers within specific sentence ranges. Notably, the largest segment represents customers with 0-10 sentences, accounting for 40.9% (70,618 customers) of the dataset, highlighting a significant concentration in this lowest range. On the right side, the bar chart provides a clear, numerical breakdown of customer counts per sentence bin, offering a different perspective from the pie chart. This bar chart confirms the trend seen in the

pie chart, showing a steep decline in customer numbers as the number of sentences increases. Both charts are instrumental in offering a dual perspective—proportional and absolute—on how customers are distributed based on the number of sentences, making the data accessible and comprehensible for strategic analysis or decision-making.

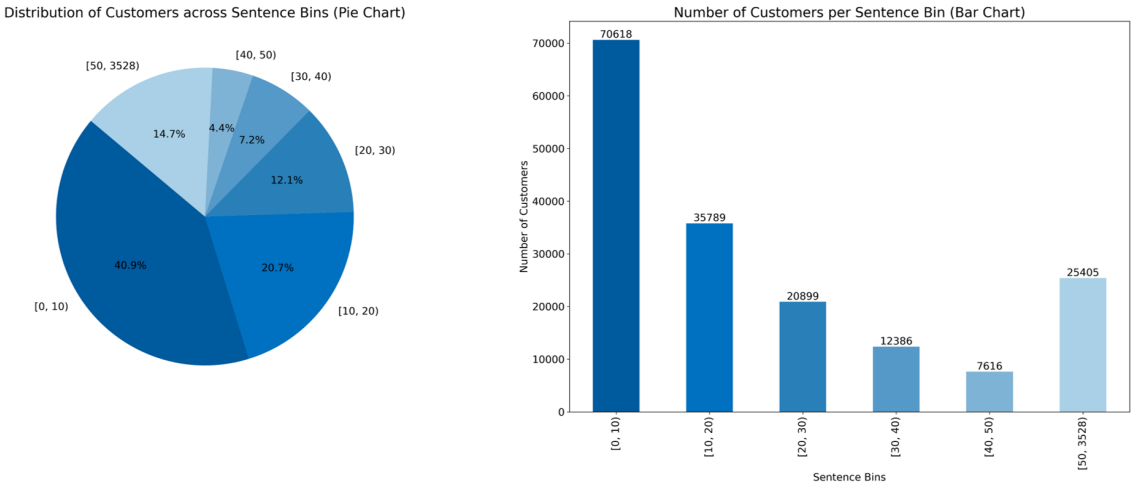


Figure 3.5: Analyzing Customer Composition and Distribution in Sentence Bins.

The data presented below (Table 3.2) show statistics regarding the length of sentences, specifically, the number of words in each sentence, categorized into different groups based on the number of sentences. An analysis of the main conclusions derived from these statistics is provided afterwards:

- **Count:** This indicates the number of sentences falling into each bin.
- **Mean:** The average length of sentences within each bin.
- **Std:** Standard deviation of sentence lengths within each bin, providing insight into the dispersion or spread of data.
- **Min:** The minimum length of sentences within each bin.
- **25%, 50%, 75%:** These percentiles represent the quartiles of the sentence length distribution, highlighting the spread of data within each bin.
- **Max:** The maximum length of sentences within each bin.

Sentences_Bins	count	mean	std	min	25%	50%	75%	max
[0, 10)	70,618	21.84	21.46	1.00	6.0	15.0	30.0	323.0
[10, 20)	35,789	71.02	45.77	10.0	36.0	59.0	97.0	539.0
[20, 30)	20,899	118.19	73.74	20.0	62.0	102.0	162.0	1292.0
[30, 40)	12,386	171.04	100.39	30.0	93.0	153.0	234.0	960.0
[40, 50)	7,616	222.56	126.57	40.0	128.0	200.0	298.0	1309.0
[50, 3528)	25,405	478.79	435.10	50.0	221.0	360.0	585.0	7641.0

Table 3.2: Descriptive Statistics for Customer Sentence Bins.

By examining these statistical data, we can obtain valuable information regarding the distribution and characteristics of sentence lengths within each bin. Our conclusions are described below:

- **[0, 10):** Sentences in this category are short, with an average length of about 22 words. The standard deviation is relatively low (21.46), indicating that most data points are clustered around the mean. This category has the highest number of sentences (70,618), suggesting that most texts include many short sentences.

- **[10, 20):** In this category, the average sentence length increases to 71 words, with a standard deviation of 45.77, showing greater variation in sentence length.

- **[20, 30), [30, 40), [40, 50):** As the range of sentences increases, so does the average length. The standard deviation also increases, indicating there are larger deviations in sentence length as we move to longer categories. Especially the category [50, 3528) shows immense variation with a standard deviation of 435, indicating the presence of extremely lengthy sentences.

- **[50, 3528):** This category includes the longest sentences with the highest maximum value (7,641 words). While the data may be rarer, some sentences are exceptionally lengthy.

The overarching conclusion is that as we examine broader ranges of word counts, the average sentence length increases, and the variation in these lengths also increases, indicating greater diversity in writing styles and the structure of longer texts.

3.2 FNS Dataset

In our continuous efforts to enhance and rigorously evaluate our summarization models on more comprehensive datasets, we utilize financial narratives from the Financial Narrative Summarisation (FNS) 2023 dataset, which includes documents in three languages: Greek, English, and Spanish. These documents were chosen for their distinctive features and the unique challenges they present, attributes often absent from more generic financial datasets.

The **Financial Narrative Summarisation (FNS) 2023** [24] initiative was a critical step forward in the domain of automatic text summarization applied to financial documents. It focused on generating concise, informative summaries from lengthy and intricate financial reports, particularly those of companies listed on the UK, Spain, and Greece stock exchanges. This dataset is particularly valuable because it includes financial narrative disclosures extracted from annual reports published in PDF format, providing a rich resource for training and testing summarization models.

The FNS 2023 dataset offers a wealth of unstructured financial information, blending textual narratives with data tables. The inclusion of documents in English, Spanish, and Greek—languages with varying grammatical structures and financial reporting norms—presents an excellent testbed for evaluating the adaptability and robustness of summarization models. For instance:

- **English Reports:** Typically well-structured with clear sections, often adhering to specific regulatory formats.
- **Spanish Reports:** While structured, they often feature narrative styles that differ from the direct and formal tone of English reports, requiring models to adapt to a more descriptive and sometimes context-heavy presentation.
- **Greek Reports:** These reports tend to be less structured and more varied in format, often blending financial data with narratives that are less segmented, posing a unique challenge for text summarization.

The table 3.3 provides the number of samples used for the summaries of the FNS dataset and the average word count of the financial reports in each language, further highlighting the variability in document length and complexity across the dataset.

Dataset	Number of Samples	Average Number of Words per Sample
Greek	49	37,454.265
English	549	73,178.299
Spanish	49	30,052.653

Table 3.3: Statistics for the FNS Dataset.

The FNS 2023 Shared Task provides a unique opportunity to explore the intersection of financial analysis and advanced computational methods. It challenges researchers to push the boundaries of what automated summarization systems can achieve in real-world applications, particularly in the financial domain where precision and clarity are paramount.

Our research initially concentrated on the Greek segment of the dataset (Fig. 3.6), provided by the FNS 2023 task. The decision to start with Greek financial reports was driven by their unique characteristics, such as the less standardized format compared to UK and Spanish reports. This focus allowed us to refine our summarization techniques to better handle the specific challenges presented by Greek financial documentation, including dealing with complex sentence structures and the integration of narrative and numerical data within less-defined sections.

Figure 3.6: Examples showcasing records from the FNS dataset.

To further evaluate our model's robustness and adaptability, we expanded our analysis to include the English and Spanish segments of the FNS dataset. This expansion was critical in testing our model's performance across different linguistic and regulatory environments, as financial discourse can vary significantly between these languages not only in style but also in the way financial information is reported and narrated.

Incorporating these languages into our dataset provided a comprehensive framework for assessing the model's ability to navigate and summarize financial documents

that differ in structural complexity and narrative style. This holistic approach significantly enhanced the model's ability to accurately capture and interpret financial narratives across multiple languages, thereby improving its utility for stakeholders who operate within diverse linguistic and cultural frameworks.

As a result, our model now demonstrates increased sensitivity to the specific informational needs and contexts of financial discourse in Greek, English, and Spanish. This expansion has not only refined our analysis but has also enhanced the model's generalizability and effectiveness in summarizing complex financial documents across different linguistic landscapes.

3.2.1 Pre-processing

Text pre-processing is an essential foundational step in natural language processing (NLP) and text summarization projects. This critical process involves refining raw textual data to optimize it for more detailed analysis and modeling. In our study, similar to our earlier efforts with the Qualco dataset, we employed a range of pre-processing techniques within the context of the FNS Dataset to enhance the quality and utility of the textual data. These steps were crucial for mitigating inherent noise in the dataset. Our key pre-processing activities included:

1. **Remove punctuation characters**

   Removes punctuation except for the period to simplify text and improve processing uniformity for tasks.

2. **Remove extra spaces**

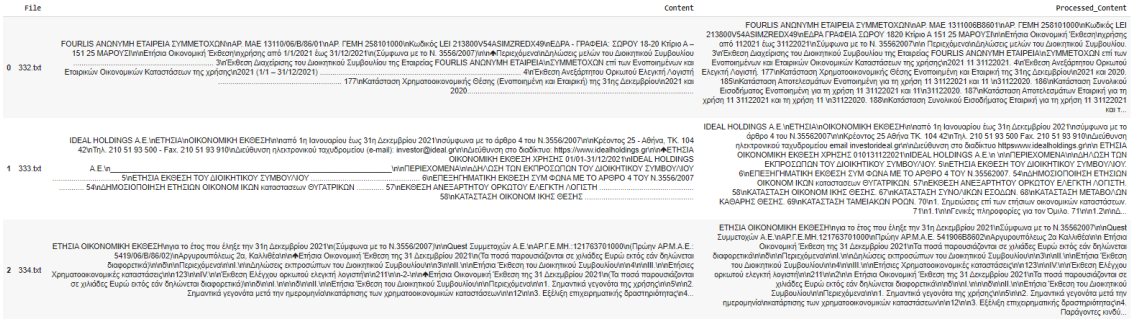
   Cleans up the data by eliminating superfluous spaces that could lead to processing errors.

3. **Remove multiple periods**

   Cleans up the text by replacing multiple consecutive periods with a single period to ensure consistency and readability.

By meticulously implementing these steps, we were able to produce a cleaner, more organized textual representation, which provided a solid foundation for our subsequent extraction of valuable insights and patterns using sophisticated analytical methods. In Fig. 3.6 shows an example featuring the original text and the processed

text in the Greek Language.



	File	Content	Processed_Content
0	332.txt	FOURLIS ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ ΣΥΜΜΕΤΟΧΩΝ\n\nΑΡ. ΜΑΕ 13110/06/Β/86/01\n\nΑΡ. ΓΕΜΗ 258101000\n\nΚωδικός LEI 213800V54ASIMZREDX49\n\nΕΔΡΑ - ΓΡΑΦΕΙΑ: ΣΟΡΟΥ 18-20 Κτίριο Α – 151 25 ΜΑΡΟΥΣΙ\n\nΕτήσια Οικονομική Έκθεση\n\nγιατη χρήση από 1/1/2021 έως 31/12/2021\n\nΣύμφωνα με το Ν. 3556/2007\n\n•Περιεχόμενα\n\nΔιάταξη μελών του Διοικητικού Συμβουλίου ..... 3\n\nΈκθεση Διαχείρισης του Διοικητικού Συμβουλίου της Εταιρείας FOURLIS ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ\n\nΣΥΜΜΕΤΟΧΩΝ επί των Ενοποιημένων και Εταιρικών Οικονομικών Καταστάσεων της χρήσης\n\n2021 (1/1 – 31/12/2021) ..... 4\n\nΈκθεση Ανεξάρτητου Ορκωτού Ελεγκτή Λογιστή ..... 177\n\nΚατάσταση Χρηματοοικονομικής Θέσης (Ενοποιημένη και Εταιρική) της 31ης Δεκεμβρίου\n\n2021 και 2020	FOURLIS ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ ΣΥΜΜΕΤΟΧΩΝ\n\nΑΡ. ΜΑΕ 13110/06/Β/86/01\n\nΑΡ. ΓΕΜΗ 258101000\n\nΚωδικός LEI 213800V54ASIMZREDX49\n\nΕΔΡΑ ΓΡΑΦΕΙΑ ΣΟΡΟΥ 18\n\n20 Κτίριο Α 151 25 ΜΑΡΟΥΣΙ\n\nΕτήσια Οικονομική Έκθεση\n\nγιατη χρήση από 1/1/2021 έως 31/12/2021\n\nΣύμφωνα με το Ν. 3556/2007\n\nΠεριεχόμενα\n\nΔιάταξη μελών του Διοικητικού Συμβουλίου ..... 3\n\nΈκθεση Διαχείρισης του Διοικητικού Συμβουλίου της Εταιρείας FOURLIS ΑΝΩΝΥΜΗ ΕΤΑΙΡΕΙΑ\n\nΣΥΜΜΕΤΟΧΩΝ επί των Ενοποιημένων και Εταιρικών Οικονομικών Καταστάσεων της χρήσης\n\n2021 11 31122021. 4\n\nΈκθεση Ανεξάρτητου Ορκωτού Ελεγκτή Λογιστή. 177\n\nΚατάσταση Χρηματοοικονομικής Θέσης Ενοποιημένη και Εταιρική της 31ης Δεκεμβρίου\n\n2021 και 2020. 185\n\nΚατάσταση Αποτελεσμάτων Ενοποιημένη για τη χρήση 11 31122021 και 11 \n\n31122020. 186\n\nΚατάσταση Συνολικού Εισοδήματος Ενοποιημένη για τη χρήση 11 31122021 και 11\n\n31122020. 187\n\nΚατάσταση Αποτελεσμάτων Εταιρική για τη χρήση 11 31122021 και τη χρήση 11 \n\n31122020. 188\n\nΚατάσταση Συνολικού Εισοδήματος Εταιρική για τη χρήση 11 31122021 και τ...
1	333.txt	IDEAL HOLDINGS Α.Ε.\n\nΕΤΗΣΙΑ\n\nΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ\n\nκατά τη λογουσρέου έως 31η Δεκεμβρίου 2021\n\nσύμφωνα με το άρθρο 4 του Ν.3556/2007\n\nκΚρέοντος 25 - Αθήνα, ΤΚ. 104 42\n\nΤηλ. 210 51 93 500 - Fax: 210 51 93 910\n\nΔιεύθυνση ηλεκτρονικού ταχυδρομείου (e-mail): investor@ideal.gr\n\nΔιεύθυνση στο διαδίκτυο: https://www.idealholdings.gr/\n\n•ΕΤΗΣΙΑ ΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ ΧΡΗΣΗΣ 01/01-31/12/2021\n\nIDEAL HOLDINGS Α.Ε.\n\n\n\nΠΕΡΙΕΧΟΜΕΝΑ\n\nΔΙΑΛΟΣΗ ΤΩΝ ΕΚΠΡΟΣΩΠΩΝ ΤΟΥ ΔΙΟΙΚΗΤΙΚΟΥ ΣΥΜΒΟΥΛΙΟΥ ..... 6\n\nΕΠΕΞΗΓΗΜΑΤΙΚΗ ΕΚΘΕΣΗ ΣΥΜ-ΦΩΝΑ ΜΕ ΤΟ ΑΡΘΡΟ 4 ΤΟΥ Ν.3556/2007 ..... 54\n\nΔΗΜΟΣΙΟΠΟΙΗΣΗ ΕΤΗΣΙΩΝ ΟΙΚΟΝΟΜ ΙΚΩΝ καταστασεων ΘΥΓΑΤΡΙΚΩΝ ..... 57\n\nΕΚΘΕΣΗ ΑΝΕΞΑΡΤΗΤΟΥ ΟΡΚΩΤΟΥ Ε\n\nΛΕΓΚΤΗ /ΛΟΓΙΣΤΗ ..... 58\n\nΚΑΤΑΣΤΑΣΗ ΟΙΚΟΝΟΜ ΙΚΗΣ ΘΕΣΗΣ	IDEAL HOLDINGS Α.Ε.\n\nΕΤΗΣΙΑ\n\nΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ\n\nκατά τη λογουσρέου έως 31η Δεκεμβρίου 2021\n\nσύμφωνα με το άρθρο 4 του Ν.3556/2007\n\nκΚρέοντος 25 Αθήνα ΤΚ. 104 42\n\nΤηλ. 210 51 93 500 Fax: 210 51 93 910\n\nΔιεύθυνση ηλεκτρονικού ταχυδρομείου email investor@ideal.gr\n\nΔιεύθυνση στο διαδίκτυο https\n\nwww.idealholdings.gr\n\nΕΤΗΣΙΑ ΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ ΧΡΗΣΗΣ 01013112021\n\nIDEAL HOLDINGS Α.Ε.\n\n\n\nΠΕΡΙΕΧΟΜΕΝΑ\n\nΔΙΑΛΟΣΗ ΤΩΝ ΕΚΠΡΟΣΩΠΩΝ ΤΟΥ ΔΙΟΙΚΗΤΙΚΟΥ ΣΥΜΒΟΥΛΙΟΥ ..... 6\n\nΕΠΕΞΗΓΗΜΑΤΙΚΗ ΕΚΘΕΣΗ ΣΥΜ ΦΩΝΑ ΜΕ ΤΟ ΑΡΘΡΟ 4 ΤΟΥ Ν.35562007. 54\n\nΔΗΜΟΣΙΟΠΟΙΗΣΗ ΕΤΗΣΙΩΝ ΟΙΚΟΝΟΜ ΙΚΩΝ καταστασεων ΘΥΓΑΤΡΙΚΩΝ. 57\n\nΕΚΘΕΣΗ ΑΝΕΞΑΡΤΗΤΟΥ ΟΡΚΩΤΟΥ Ε\n\nΛΕΓΚΤΗ /ΛΟΓΙΣΤΗ. 58\n\nΚΑΤΑΣΤΑΣΗ ΟΙΚΟΝΟΜ ΙΚΗΣ ΘΕΣΗΣ. 67\n\nΚΑΤΑΣΤΑΣΗ ΣΥΝΟΛΙΚΩΝ ΕΣΟΔΩΝ. 68\n\nΚΑΤΑΣΤΑΣΗ ΜΕΤΑΒΟΛΩΝ ΚΑΘΑΡΗΣ ΘΕΣΗΣ. 69\n\nΚΑΤΑΣΤΑΣΗ ΤΑΜΕΙΑΚΩΝ ΡΟΩΝ. 70\n\n1. Σημειώσεις επί των ετήσιων οικονομικών καταστάσεων. 7\n\n1.1\n\nΓενικές πληροφορίες για τον Όμιλο. 7\n\n1.2\n\nΠ...
2	334.txt	ΕΤΗΣΙΑ ΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ\n\nγια το έτος που έληξε την 31η Δεκεμβρίου 2021\n\nΣύμφωνα με το Ν.3556/2007\n\nQuest Συμμετοχών Α.Ε.\n\nΑΡ.Γ.Ε.ΜΗ.: 121763701000\n\nΠρώην ΑΡ.Μ.Α.Ε.: 5419/06/Β/86/02/η\n\nΑγγυροπτάλευς 2α, Καλλιθέα\n\n•Ετήσια Οικονομική Έκθεση της 31 Δεκεμβρίου 2021\n\nΤα ποσά παρουσιάζονται σε χιλιάδες Ευρώ εκτός εάν δηλώνεται διαφορετικά\n\nΠεριεχόμενα\n\nΔιάλυσης εκπροσώπων του Διοικητικού Συμβουλίου\n\nΕπιστολή Έκθεση του Διοικητικού Συμβουλίου\n\nΕπίσης Χρηματοοικονομικές καταστάσεις\n\n123\n\nV\n\nΈκθεση Ελέγχου ορκωτού ελεγκτή λογιστή\n\n211\n\n2-\n\nΕτήσια Οικονομική Έκθεση της 31 Δεκεμβρίου 2021\n\nΤα ποσά παρουσιάζονται σε χιλιάδες Ευρώ εκτός εάν δηλώνεται διαφορετικά\n\nΈκθεση του Διοικητικού Συμβουλίου\n\nΠεριεχόμενα\n\n1. Σημαντικά γεγονότα της χρήσης\n\n2. Σημαντικά γεγονότα μετά την ημερομηνία\n\nκατάρτισης των χρηματοοικονομικών καταστάσεων\n\n12\n\n3. Εξέλιξη επιχειρηματικής δραστηριότητας\n\n4...	ΕΤΗΣΙΑ ΟΙΚΟΝΟΜΙΚΗ ΕΚΘΕΣΗ\n\nγια το έτος που έληξε την 31η Δεκεμβρίου 2021\n\nΣύμφωνα με το Ν.3556/2007\n\nQuest Συμμετοχών Α.Ε.\n\nΑΡ.Γ.Ε.ΜΗ.:121763701000\n\nΠρώην ΑΡ.Μ.Α.Ε. 54190\n\nΒ\n\n86\n\n02\n\nΑγγυροπτάλευς 2α Καλλιθέα\n\nΕτήσια Οικονομική Έκθεση της 31 Δεκεμβρίου 2021\n\nΤα ποσά παρουσιάζονται σε χιλιάδες Ευρώ εκτός εάν δηλώνεται διαφορετικά\n\nΠεριεχόμενα\n\nΔιάλυσης εκπροσώπων του Διοικητικού Συμβουλίου\n\nΕπιστολή Έκθεση του Διοικητικού Συμβουλίου\n\nΕτήσιες Χρηματοοικονομικές καταστάσεις\n\n123\n\nV\n\nΈκθεση Ελέγχου ορκωτού ελεγκτή λογιστή\n\n211\n\nΕτήσια Οικονομική Έκθεση της 31 Δεκεμβρίου 2021\n\nΤα ποσά παρουσιάζονται σε χιλιάδες Ευρώ εκτός εάν δηλώνεται διαφορετικά\n\nΈκθεση του Διοικητικού Συμβουλίου\n\nΠεριεχόμενα\n\n1. Σημαντικά γεγονότα της χρήσης\n\n2. Σημαντικά γεγονότα μετά την ημερομηνία\n\nκατάρτισης των χρηματοοικονομικών καταστάσεων\n\n12\n\n3. Εξέλιξη επιχειρηματικής δραστηριότητας\n\n4. Παράγοντες κινδύ...

Figure 3.7: Example of Text Pre-processing: Original and Processed.

In conclusion, the use of both the QUALCO dataset and the Financial Narrative Summarisation (FNS) datasets has been instrumental in advancing our research on automated text summarization. The comprehensive and detailed QUALCO data provided a robust foundation for developing our models, while the diverse and multilingual FNS datasets enabled thorough evaluation and refinement. This combination allowed us to address unique challenges across different financial contexts and languages, as we tested our methods on both the specific agent-based discussions and the more real-world related multilingual FNS dataset for benchmarking purposes, ultimately enhancing the robustness and applicability of our summarization techniques.

Chapter 4

Method

This section presents the architecture of our system for automated text summarization using various techniques. Specifically, to fully leverage the data in our dataset, we adopted an innovative approach by calculating sentence weights based on the publication date of each comment. Additionally, to ensure the clarity and precision of our summaries, we retained unique sentences from the comments, prioritizing those with the highest weights in cases of repetition. Furthermore, to create summaries with or without applying weights, we employed three techniques: K-Means cluster analysis, TF-IDF, and random sentence selection. However, for the FNS dataset, we exclusively used the K-Means cluster analysis method to handle its unique structure and complexity. This chapter aims to explore the effectiveness of these methods in text restructuring and their impact on the management and presentation of information.

4.1 Weights Calculation

This section outlines the method employed to assign weights to customer comments based on their chronological order. Utilizing the dates provided by Qualco, we apply a time-decay weighting scheme that reflects the diminishing impact of past events, mirroring the natural fading of human memory.

Customer feedback is a vital component of business analysis, but not all feedback holds the same value over time. To adjust for the relevance of older comments, we use an exponential decay function [25][26], widely used in fields such as signal processing and econometrics for modeling time-based fading processes. This function helps decrease the importance of events exponentially as they recede into the

past. The key element of our weighting strategy is the application of a decay factor that modifies the significance of each comment based on its age relative to the last recorded comment for a customer. The following steps were taken to implement this methodology:

1. **Data Preparation:** The first step involves preparing the data by converting the dates from strings to a format suitable for calculations. This conversion is critical as it allows for precise time-based calculations. Each comment's date is transformed into a DateTime object, a standard Python format for handling dates and times, which supports arithmetic operations needed for subsequent calculations.

2. **Identification of Last Comment Date:** To adjust the weights of comments based on their timeliness, it is necessary to identify the most recent comment for each customer. This involves scanning the dataset to determine the last date any given customer left a comment. This date becomes a reference point (current_date) against which all other comments from the same customer are compared.

3. **Weight Calculation Using Exponential Decay:**

   - **Decay Function Application:** The core of our weighting strategy is the exponential decay function $e^{-decay\_parameter*days\_ago}$. This function is applied to calculate the weight of each comment based on how many days have passed since it was made relative to the most recent comment of the same customer. The formula for this function exponentially decreases the weight of older comments, mimicking the natural decline in relevance over time. The choice of function is a mathematically sound way to model time decay, and it is commonly used in various fields like finance [27], recommendation systems, and data analysis when dealing with time-sensitive data.

   - **Setting the Decay Parameter:** The decay parameter is used to control the rate at which comment relevance declines. A typical value might be set such that the importance of feedback declines by a certain percentage

each day. Adjusting this parameter allows for fine-tuning the sensitivity of the analysis to past data. In the present analysis, the parameter was set by default to 0.05, representing the assumption that the relevance of comments decreases by about 5% each day. The choice of decay parameter was made because a 1% decay rate is suitable for analyses where long-term feedback is vital. In contrast, a 10% decay rate is useful for scenarios where the most recent feedback is critical and older comments quickly lose value. Therefore, the 5% decay rate prioritizes recent comments while still allowing older feedback to significantly contribute to the analysis.

By employing a 5% decay parameter, the analysis provides a balanced understanding of how comment relevance declines over time, affecting the overall insights derived from the feedback. This balanced decay rate meets the specific requirements of our analysis by maintaining a comprehensive view of customer feedback over time.

4. **Application of Weights to Comments:** Each comment in the dataset is processed individually to apply the calculated weight. This involves:
   - Retrieving the last action date for the customer associated with each comment.
   - Using the decay function to compute the weight based on the time elapsed since the comment was made.
   - Storing the weighted value alongside other relevant information from the comment, such as the customer ID, the text of the comment, and both the original and last action dates.
5. **Compilation of Results:** After processing all the comments and calculating their weighting, the results are compiled in a structured table. This compilation includes all relevant details necessary for further analysis or reporting, ensuring that the impact of each comment is appropriately represented according to its timeliness.

This weight calculation method reflects the importance of the most recent customer

comments, aiming to support decision-making based on up-to-date feedback. The method outlined above is essential for analyzing customer comments in a manner that realistically captures their diminishing relevance over time. By prioritizing recent feedback, this method enhances our ability to derive meaningful insights and make well-informed decisions that are reflective of current customer sentiments.

4.2 Unique Sentences

Upon observing the data post-preprocessing and grouping the customer comments from the QUALCO dataset, it became evident that repetitions of similar sentences occurred across various customers' entries. To address this redundancy and maintain the uniqueness of the dataset, we implemented a strategy to retain unique comments for each customer. Specifically, we preserved those with the highest assigned weights for duplicated comments (the data are stored in a new "Higher_Weights" column). This decision aligns with the weighting approach previously detailed, where comments closer to the implementation date are considered more significant and thus assigned a bigger weight.

Furthermore, we focused on unique sentences to ensure meaningful clustering in our analysis using K-Means with TF-IDF. Duplicates can introduce bias in the clustering process, leading to skewed results and redundant clusters. By removing duplicates, we reduced computational load and enhanced the interpretability of the clusters, making each one represent a distinct theme or topic without being overwhelmed by repetitions.

In Table 4.1 below, we provide a detailed example to demonstrate our approach to handling data redundancy. This example showcases a scenario involving a customer who posted multiple similar comments. We retained only the distinct sentences, accompanied by their respective weights, to enhance the clarity and accuracy of our analysis. Additionally, this ensured that our K-Means clustering with TF-IDF was based on a balanced dataset, free from the undue influence of duplicated comments. The removal of duplicates before clustering allows for more accurate and efficient computation, ultimately leading to more meaningful insights from the customer feedback data.

Processed Comment	Weight	Unique Sentences	Higher Weights
κα.	[0.00013638892648201,	*κα.*	[0.00013638892648201,
δεσ σημειωσεισ.	0.00174674713626112,	*δεσ σημειωσεισ.*	1.0,
κατκχωρη actioncode λιστα.	0.01290681258047986,	*κατκχωρη actioncode λιστα.*	0.02024191144580438,
κατκχωρη actioncode λιστα.	0.02024191144580438,	*καταχ δεσ σημειωσεισ.*	0.03688316740123999,
καταχ δεσ σημειωσεισ.	0.03688316740123999,	*αναζητη δεν υπαρουν τηεφωνα επικοινωνιασ.*	0.000000064927147715]
δεσ σημειωσεισ.	0.6376281516217733,		
δεσ σημειωσεισ.	0.0450492023935578,		
δεσ σημειωσεισ.	1.0,		
αναζητη δεν υπαρουν τηεφωνα επικοινωνιασ.	0.000000064927147715]		

Table 4.1: Unique Sentences Calculation.

This methodology was specifically designed to identify recurring patterns and prioritize comments based on their historical significance. By ensuring that each comment remains unique to each customer and assigning weights accordingly, we aimed to enhance the analytical precision and deliver deeper, more actionable insights. It is important to note that the distribution of comments per customer remains consistent throughout this process. Figure 4.1 provides a visual representation of this distribution, illustrating how our weighting methodology impacts the analyzed data without altering the original frequency of comments per customer.

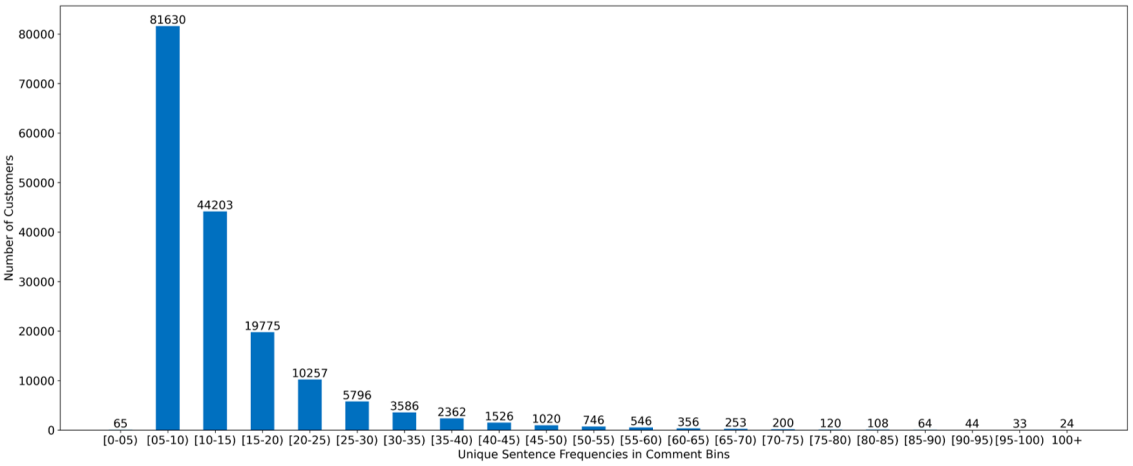


Figure 4.1: Number of Customers per Comment Bin (Unique Sentences).

It is apparent that after the application of our methodology, the number of customers in the initial bins, which contain the majority of customers, has increased. This indicates that many customers had repetitive statements in their comments. From this observation, we conclude that our methodology for ensuring the unique-

ness of comments is effective, improving the accuracy and quality of our analysis without distorting the original distribution of comments.

4.3 Techniques

Building on the insights gleaned from the previous subsection, our analysis revealed that a majority of customers predominantly fall into the bins containing [5-10), [10-15) and [15-20) comments based on the QUALCO data. This distribution provides a critical insight into the volume of feedback per customer, which is instrumental in tailoring our text summarization efforts effectively.

To address the varying needs of information condensation required by different customer comment volumes, we have decided to implement a tiered approach for our summary generation system. This approach ensures that the summary generated is proportional to the volume of feedback provided by each customer, enhancing the relevance and usability of the summaries. Based on the distribution of comments across our customer base, we segmented the lengths of the summaries into four main categories: 5, 10, 15, and 20 sentences. This gradation allowed us to maintain a balance between completeness and brevity, tailored to the volume of input data. With the summary lengths established to align with the comment frequencies identified, we now delve into the specific summarization techniques employed to craft these tailored summaries. Figure 4.2 illustrates the flow that was followed, based on the QUALCO dataset, and we analyze each method in the following sections.

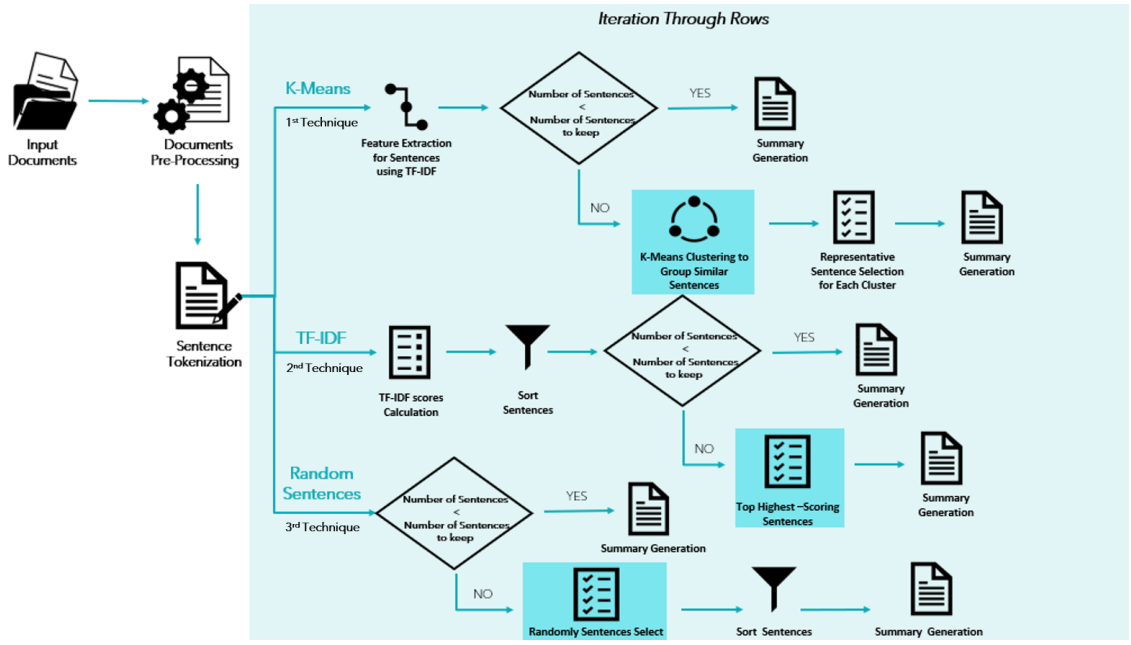


Figure 4.2: System Architecture with or without Weights.

In summary, The K-means algorithm clusters sentences and selects representative sentences from each cluster to create a summary. Sentences are vectorized using TF-IDF, and K-means clustering is applied to these vectors. The summary is generated by selecting sentences closest to the cluster centroids, ensuring diverse content representation. The TF-IDF method assigns a weighted importance to each sentence based on its term frequency-inverse document frequency score. Sentences are ranked according to their weighted TF-IDF scores, and the top-scoring sentences are selected for the summary, emphasizing sentences with unique and significant terms. The Random method randomly selects a specified number of unique sentences to create a summary, ensuring that the selected sentences are unique and in their original order, providing a quick and unbiased summary without any weighting or clustering. Each of these methods ensures that summaries are effectively generated, aligning with the quantity and nature of customer feedback. This multi-faceted approach provides flexibility and efficiency in handling varying comment volumes, and these techniques incorporate weights to refine the summary generation process, with the code affecting the areas shown in dark shading in the figure.

4.3.1 K-Means

K-means is the first of three techniques used to define a flow for producing summaries. This technique, along with the others, can be applied with or without weight calculations, as we will see in the following paragraphs. Importantly, K-means is the only method that was also used with the FNS dataset, demonstrating its versatility and effectiveness across different datasets. This application will be analyzed in detail in the subsection 4.3.1.3.

4.3.1.1 Without the Effect of Weights

In this context, K-Means analysis is utilized to cluster similar sentences and select representative statements that encapsulate key themes within the comments. This clustering method is considered highly effective in text summarization because it organizes sentences into distinct groups based on similarity, which allows for the identification and selection of the most representative sentences from each cluster. The main advantage of this approach is its potential to reduce redundancy while

preserving the essence of the feedback, aiming to make the summary both concise and informative. The steps to create the summary are as follows (Fig. 4.2):

1. **Preparation of Sentence Vectors:** The process begins by extracting feature vectors for each sentence using a TF-IDF vectorizer. This step transforms the text data into a numerical format that K-Means can process, emphasizing the importance of unique words in representing the content of each sentence.
2. **Clustering Sentences:** With the sentences converted into vectors, K-Means clustering is applied. The number of clusters is set to the desired number of sentences in the summary (5,10,15,20 sentences). This ensures that each cluster ideally represents a unique aspect of the input text.
3. **Selection of Representative Sentences:**
   - For each cluster, the sentence closest to the cluster center (centroid) is selected as the most representative of that group. This step involves calculating the distance of each sentence from the center of its cluster and selecting the sentence with the minimum distance, which effectively captures the central theme of the cluster.
   - A tolerance threshold is used to identify sentences that are very close to being the most central, providing a buffer against minor variations in text that might otherwise lead to the selection of a less representative sentence. This threshold ensures that the selection process accounts for slight differences in sentence placement relative to the cluster center, enhancing the summary's representativeness. The code employs K-Means clustering to group sentences, selecting the closest sentence to each cluster center as the "central" sentence. However, due to textual variations, multiple sentences may be nearly equidistant from the center. By applying a tolerance threshold, the code includes all sentences within a small distance from the central sentence, ensuring that the most representative sentences are chosen for the summary. This method enhances the robustness and accuracy of the generated summary by considering all relevant sentences within the specified tolerance.
4. **Integration into Data Framework:** Once the representative sentences are

selected from each cluster, they are combined to form the summary. The sentences are sorted in their original order to maintain a logical flow in the resulting summary.

5. **Construction of Summary:** The generated summary is then stored back into the dataset under a new column specifically designated for these summaries. This step finalizes the summarization process for each entry in the dataset.

This implementation effectively addresses the challenge of summarizing extensive customer feedback by focusing on reducing redundancy without losing valuable information. The use of K-Means aims to guide the summarization process to include sentences that reflect the diverse viewpoints and information presented by the customers. The summaries generated through this method can be used for subsequent analysis, providing a quick yet comprehensive overview of each customer's feedback.

4.3.1.2 With the Effect of Weights

This approach integrates the textual significance evaluation of TF-IDF with the clustering capability of K-Means, further enhanced by incorporating sentence weights. This method ensures that the summaries not only encapsulate the most significant textual themes but also align with weighted priorities, such as the urgency or impact of the feedback.

1. **Preparation and Vectorization:** Sentences are extracted and cleaned to ensure meaningful input into the TF-IDF vectorizer, which then converts these sentences into a numerical format that represents their textual significance.

2. **Incorporation and Application of Weights:** Weights, first calculated based on a decay function, are applied to the TF-IDF scores, adjusting the resultant vectors to reflect both the inherent importance of the words in the sentences and their external significance as determined by the weights. This creates a combined score that influences both the content and the priority of each sentence.

3. **Clustering with K-Means:** The adjusted scores are then clustered using K-Means, which groups sentences based on the similarity of their weighted TF-IDF scores. This method identifies patterns within the data and helps

ensure that the summaries represent the themes of each cluster, highlighting the most relevant and important sentences within each group.

4. **Selection of Representative Sentences:** For each cluster, the sentence closest to the cluster center — considered the most representative of that cluster — is selected. This is determined by finding the sentence within each cluster that has the shortest distance to the cluster center, indicating it best encapsulates the cluster's theme.
5. **Summary Construction:** The selected sentences are sorted according to their original order in the text to maintain narrative coherence. They are then concatenated to form the summary, ensuring that it is readable and contextually accurate.
6. **Integration into Data Framework:** The final summary is stored in a newly created column within the dataset, allowing for easy access and further analysis. This integration ensures that the summaries can be readily compared with other data or used in subsequent decision-making processes.

This updated summarization technique enhances the relevance and utility of the summaries by ensuring they are both representative of the textual content and aligned with specific business or analytical priorities. It is particularly beneficial in scenarios where the importance of comments varies, allowing stakeholders to quickly identify and focus on key issues and sentiments. By combining TF-IDF and K-Means with weights, the summaries are tailored to reflect the most pertinent information, providing a focused overview of customer feedback that is crucial for informed decision-making.

4.3.1.3 K-means in the FNS dataset

In this implementation, K-Means clustering is employed to effectively summarize extensive content by condensing it into a target word count of 1,000 words. The purpose of using this clustering technique is to segment the text into coherent groups that are thematically similar, making it easier to identify and select key sentences that collectively convey the core messages while aiming to adhere to the word count limit. This method is particularly adept at maintaining essential information without significant redundancy, ensuring that the summary is not only concise but also

comprehensive and reflective of the original content:

1. **Preparation of Sentence Vectors:**
    - The content is first processed to strip spaces and remove any empty sentences, ensuring that only meaningful text is analyzed. Unique sentences are identified to avoid redundancy.
    - Each sentence is then converted into numerical feature vectors using a TF-IDF vectorizer. This transformation is crucial as it allows the clustering algorithm to measure textual similarity based on the significance of terms within sentences.
2. **Clustering Sentences:** The sentences are grouped using the K-Means algorithm. The number of clusters is dynamically determined based on the total number of sentences, with an aim to create a manageable number of clusters that can logically represent different facets of the input text.
3. **Selection of Representative Sentences:**
    - Within each cluster, sentences are ordered by their proximity to the cluster center. This step is vital as it prioritizes sentences that are most representative of each cluster's thematic core.
    - Sentences are then selected sequentially from each cluster while monitoring the cumulative word count. This selective process continues until the summary reaches or exceeds the target word count of 1000 words, ensuring that the final summary is concise yet comprehensive.
4. **Integration into Data Framework:** Once the sentences are chosen, they are sorted chronologically based on their original order in the text. This helps in maintaining a logical and coherent flow in the narrative of the summary.
5. **Construction of Summary:** The selected sentences are concatenated to form the final summary. This summary is then stored in a new column within the dataset, specifically designated for these summaries. This step concludes the summarization process for each entry in the dataset.

This methodology not only addresses the challenge of distilling large volumes of text into concise summaries but also ensures that these summaries adequately represent

the diverse perspectives and key information contained in the original content. The summaries generated are intended for further analysis or as a digestible overview of complex documents.

4.3.2 TF-IDF Method

The second technique is the TF-IDF method which ranks sentences by their importance based on term frequency-inverse document frequency scores, and can also be applied with or without weight calculations. The Fig.4.2 illustrates the flow described below.

4.3.2.1 Without the Effect of Weights

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Applied to text summarization, the TF-IDF method emphasizes words that are particularly distinctive in a set of comments, helping to pinpoint significant phrases and sentences crucial for crafting a meaningful and content-rich summary. Below we analyze the methodology used for the specific technique:

1. **Sentence Vectorization:** The process begins by converting the text data into a format suitable for analysis using a TF-IDF vectorizer. This step involves calculating the TF-IDF scores for each word in each sentence, effectively transforming the sentences into a vector space based on their textual content.
2. **Sentence Importance Evaluation:** After vectorization, each sentence is assigned a score that represents its overall significance within the entire set of comments. This score is determined by summing the TF-IDF values of all words in the sentence, giving higher scores to sentences containing words that are rare across other comments but frequent within their own context.
3. **Sentence Selection for Summary:**
    - With scores assigned, the sentences are ranked according to their TF-IDF scores, and the top sentences—those with the highest scores—are selected for the summary. The number of sentences selected is predeter-

mined (5,10,15,20 sentences), ensuring that the summary is concise yet comprehensive.

- To maintain the logical flow of the original text, the selected sentences are sorted back into their original order before being combined to form the summary.

4. **Construction of Summary:** The summary is constructed by concatenating the selected sentences, their order of appearance is maintained as in the original text. This ensures that the most informative parts of the comments are included in the summary, providing an accurate and insightful snapshot.

5. **Integration into Data Framework:** Finally, the constructed summary is stored in a designated column within the dataset. This allows for easy access and review, integrating the summary directly with the related data for further analysis or presentation.

This implementation of the TF-IDF method in summarization effectively highlights the most informative sentences from customer comments. By focusing on sentences that contain distinctive words, the method ensures that the summary captures the essence of the feedback without unnecessary redundancy. This alternative flow serves as a baseline for comparison reasons against the k-means clustering approach followed in the previous section.

4.3.2.2 With the Effect of Weights

In this approach, we upgrade the traditional TF-IDF methodology by integrating a weighting mechanism that factors in the relative importance of each sentence as dictated by external criteria (e.g., customer feedback relevance or urgency). This method is designed to ensure that the summaries not only reflect the intrinsic textual significance of the sentences but also align with specific analytical or business priorities through the applied weights. We elaborate on the steps below:

1. **Preparation and Sentence Vectorization:** Sentences are initially extracted and cleaned of any duplicates or empty entries to ensure that each sentence is unique and contributes meaningfully to the summary. It is an additional confirmation step as the summaries are based on the unique sentences for

each customer. The TF-IDF vectorizer then transforms these sentences into a numerical format that represents their textual significance within the corpus.

2. **Incorporation of Weights:** Each sentence is associated with a specific weight, which may reflect factors such as recentness, customer sentiment strength, or any other relevant metric. These weights are used to adjust the TF-IDF scores, amplifying the influence of sentences deemed more important.

3. **Weighted Score Calculation:** The traditional TF-IDF scores and the associated weights are combined to compute a weighted score for each sentence. This is achieved by multiplying the TF-IDF scores with the corresponding weights, resulting in scores that reflect both the textual relevance and the weighted importance.

4. **Sentence Selection for Summary:** Sentences are ranked based on their weighted scores, and the top sentences - those with the highest combined scores - are selected for inclusion in the summary. This process ensures that the summary emphasizes content that is both inherently significant and strategically important.

5. **Summary Construction:** Selected sentences are sorted back into their original order and concatenated to form a coherent summary. This preserves the narrative flow and ensures that the summary is easily understandable and contextually accurate.

6. **Integration into Data Framework:** The final summary is stored in a dedicated column within the dataset, facilitating easy access for further analysis or review. This step integrates the summary directly with its corresponding data, enhancing the utility of the dataset for decision-making or reporting purposes.

Integrating weights into the TF-IDF summarization process enhances the relevance of the summaries by ensuring they reflect prioritized insights. This method is particularly beneficial when dealing with large volumes of text where key information must be quickly identified and highlighted. By adjusting the summaries according to predefined weights, the summaries become more tailored to specific needs, providing stakeholders with focused and meaningful insights into the data.

4.3.3 Random Sentence Selection

The Random method is utilized as a baseline technique in text summarization to ensure variety and capture unique aspects of the text that structured methods like K-Means and TF-IDF might miss, providing a quick and unbiased summary generation approach that can be employed independently or with weight adjustments for added refinement.

4.3.3.1 Without the Effect of Weights

By incorporating randomness, this method can bring out diverse viewpoints and unexpected insights, which are crucial for a holistic understanding of customer sentiments. It effectively complements the more deterministic methods, providing a balance between structure and spontaneity in the summarization process. The steps for implementing randomization without calculating weights are described below:

1. **Sentence Preparation:** The initial step involves preparing the text by splitting the customer comments into individual sentences. This breakdown forms the basis for sentence selection and ensures that the summarization process considers each statement as a potential candidate.

2. **Sentence Count Assessment:** The total number of sentences available for each customer is counted to determine how many sentences can potentially be included in the summary. This count helps in setting realistic expectations for summary length, particularly in cases where the comment section might be brief.

3. **Determination of Sentence Quantity:** The number of sentences (5,10,15,20 sentences) to include in the summary is specified, with a maximum set to ensure conciseness and to avoid exceeding the number of existing proposals.

4. **Random Sentence Selection:**

    - If the total number of sentences is sufficient (more than one), random selection is employed to choose sentences. This randomness ensures that the summary can include a variety of sentences, thus potentially capturing different aspects of the customer's feedback that are not highlighted by other methods.

- The selected sentences are sorted according to their original order in the text. This maintains the narrative flow and makes the summary coherent and easier to understand.

5. **Summary Construction:** The chosen sentences are concatenated to form the summary. This final composition reflects a range of ideas and sentiments, presenting a broad spectrum of the customer's feedback.

6. **Integration into Data Framework:** The newly created summary is stored in a designated column within the dataset. This integration allows for seamless access to the summaries for subsequent analysis, comparison, or presentation.

The implementation of Random Sentence Selection offers a way to ensure that each summary is unique and enriched with a diverse set of perspectives. This method is particularly useful in scenarios where customer feedback varies widely, ensuring that even less dominant themes are represented. The summaries produced by this technique provide a quick, varied glimpse into the range of customer sentiments, enhancing the understanding of the overall feedback landscape.

4.3.3.2 With the Effect of Weights

In this improved version of our Random Sentence Selection technique, we integrate the decay-based weight mechanism as before to refine the randomness by factoring in the significance of each sentence. This method extends our baseline random selection process by not only ensuring diversity in the summaries but also aligning them more closely with the weighted importance of the comments. This approach is particularly useful in emphasizing sentiments or information that are deemed more critical based on predetermined criteria, such as customer feedback intensity or recentness. The steps for applying randomization to each proposal with weighting are as follows:

1. **Preparation and Weight Assignment:** Each sentence in a customer's comment is split and prepared for processing, similar to the original random selection method. Additionally, each sentence is associated with a weight, derived from 'Higher_Weights' column, which quantifies the relevance or impact of the sentence based on our analytic criteria.

2. **Weighted Random Selection:** Instead of selecting sentences purely at random, this method uses the weights of the sentences as probabilities in the selection process. This weighted random selection ensures that sentences with higher weights have a greater chance of being chosen for the summary, thus prioritizing content that carries more significance.

3. **Summary Composition:** After selecting the sentences based on their weights, the chosen sentences are sorted back into their original order to preserve the logical flow of the narrative. The summary is then constructed by concatenating these selected sentences, ensuring that it reflects the weighted perspectives and insights from the customer feedback.

4. **Integration and Review:** The final weighted summary is stored in a dedicated column within the dataset, allowing for easy access and subsequent analysis. This integration facilitates a direct comparison between weighted and non-weighted summaries, highlighting the enhancements brought by considering the weights in the summarization process.

The weighted random sentence selection method enhances the utility of the summaries by ensuring that they not only capture a broad spectrum of sentiments but also emphasize those considered most impactful. This approach is particularly beneficial in scenarios where the importance of comments varies significantly, allowing analysts and decision-makers to quickly grasp key issues and sentiments without sifting through less pertinent information. By incorporating weights, the summaries become more aligned with strategic business needs, providing a focused and nuanced view of customer feedback.

To conclude this chapter on text summarization techniques, we have explored a spectrum of methodologies including K-Means analysis, TF-IDF-based sentence selection, and Random Sentence Selection, each tailored to condense extensive customer feedback into concise, informative summaries. These strategies, implemented in their standard forms and further refined with weighted adjustments to account for the temporal relevance of sentences, have been instrumental in developing a sophisticated framework. This framework excels in capturing the essence of customer feedback, adeptly highlighting pertinent sentiments and information, thus enabling

businesses to focus on areas of significant impact.

The introduction of weighted mechanisms into our summarization processes marks a significant advancement, enhancing the summaries' relevance and strategic alignment with business priorities. By emphasizing recent feedback and issues crucial to the organization, this approach supports nuanced decision-making. It empowers stakeholders to swiftly identify and respond to key elements of feedback that are most likely to influence business operations and customer relations, ensuring that strategic actions are both informed and timely.

Furthermore, within the context of the FNS Dataset, the strategic application of K-Means clustering demonstrates a robust approach to managing vast datasets. By clustering texts and selectively summarizing them to meet a 1000-word target, this method maintains thematic integrity and coherence. The summaries produced are not only succinct but richly representative of the dataset's overarching themes, providing stakeholders with a distilled yet comprehensive view of the textual data.

This strategic integration of clustering and summarization techniques significantly enhances the utility of the produced summaries. By delivering a clear and focused snapshot of extensive data, these methods facilitate efficient decision-making processes. They provide stakeholders with the crucial insights needed to navigate complex business landscapes, making these summarization techniques invaluable tools for any organization aiming to optimize its operations and customer engagement strategies effectively.

Chapter 5

Experimental Setup

In this section, we focus on the creation and use of Gold Summaries as a benchmark for evaluating the performance of summarization algorithms. Additionally, we discuss the evaluation methods employed, including the use of the ROUGE metric, chosen for its effectiveness in assessing the quality of generated summaries compared to Gold Summaries and other reference data.

5.1 Ground Truth (Gold Summaries)

In the context of text summarization, ground truth is crucial as it provides the definitive standard for comparing and assessing the effectiveness of summarization algorithms. Typically, this ground truth is established through what are known as "Gold Summaries",which are meticulously crafted summaries considered to embody the ideal summary of the source texts.

However, challenges arise when gold summaries are not available, especially in novel or large domains where creating human summaries for all documents is impractical. In such cases, alternative evaluation techniques are needed to assess the quality of machine-generated summaries without relying on human model summaries. Techniques such as using pseudomodels, which are system summaries chosen based on their predicted scores, can help expand the set of available model summaries and improve the reliability of evaluation results [28][29].

In our project, we created Gold Summaries to be used as ground truth for evaluating text summarization algorithms tailored to customer feedback analysis. The

process of developing these summaries is structured to ensure high fidelity and relevance:

1. **Samples Selection:** Given the vast amount of information available, we opted to focus on customers with a significant number of comments to ensure representativeness. To achieve a diverse sample for summary creation, we randomly selected 61 customers whose number of comments ranged from 85 to 90.

2. **Human Summarization:** To create the Gold summaries, we prioritized the latest sentences, considering them to be the most recent and therefore the most relevant at the time of writing. It should be noted that the summaries were created from complete sentences rather than individual words.

3. **Utilization of Ground Truth:** Once established, the Gold Summaries were used to evaluate and train automatic text summarization algorithms. The performance of the algorithms was evaluated based on how well the generated summaries matched the Gold Summaries in terms of information retention, coherence, conciseness, and readability. This evaluation helps to iterate and improve summarization algorithms, guide developments in algorithm design, and refine data processing techniques.

5.2 Evaluation

Text summary evaluation metrics are essential for assessing the effectiveness of automatic summarization systems. By comparing generated summaries with reference summaries or human assessments, these metrics provide quantitative measures of summary quality, including accuracy, coherence, readability, and informativeness. Accurate assessment measurements are crucial for the development and continuous enhancement of language models. Popular evaluation metrics for text summarization include ROUGE, BLEU, BERTScore, and METEOR[30].

This thesis, ROUGE was selected as the primary evaluation metric due to its proven correlation with human judgment and its effective measurement of content overlap between generated abstracts and reference texts. ROUGE (Recall-Oriented

Understudy for Gisting Evaluation) is a comprehensive set of metrics used to evaluate both automatic summarization and machine translation. It compares automatically produced summaries or translations against a set of reference summaries, typically human-generated.
The ROUGE toolkit includes several metrics, such as ROUGE-N, which measures n-gram overlap; ROUGE-L, which utilizes the longest common subsequence; ROUGE-S, focusing on skip-bigrams; and ROUGE-W, which considers the length of the longest common subsequence to provide higher scores to longer sequences. This variety of metrics highlights ROUGE's flexibility, ease of use, and alignment with human evaluative standards, making it a robust choice for text summarization and machine translation [20][30].

For evaluating the effectiveness of the automatic summaries produced by the models, whether incorporating weights or not, and the gold summaries, the ROUGE metric was utilized. Specifically, three sub-categories of ROUGE were applied: ROUGE-1, ROUGE-2, and ROUGE-L, which assess precision, recall, and the F1 score for unigrams, bigrams, and the longest common subsequence, respectively. These metrics—**Precision, Recall, and F1 Score**—are fundamental in evaluating information retrieval and natural language processing tasks, including text summarization. Precision measures the proportion of correctly identified relevant elements out of all elements identified by the model, while recall measures the proportion of correctly identified relevant elements out of all relevant elements that should have been identified. The F1 Score, as the harmonic mean of precision and recall, provides a balanced view of the model's performance, especially when precision and recall are at odds.

In the context of ROUGE metrics, applying precision, recall, and F1 Score to different granularities, such as unigrams (ROUGE-1), bigrams (ROUGE-2), and the longest common subsequence (ROUGE-L), allows for a detailed evaluation of summarization quality. This approach considers both the accuracy and completeness of the generated summaries, which is crucial for understanding the effectiveness of summarization models and guiding further improvements[23].

For the FNS dataset specifically, an additional metric, ROUGE-SU4, was uti-

lized. ROUGE-SU4, which measures skip-bigram overlap and is particularly useful for capturing semantic similarities that are not necessarily captured by standard n-gram metrics, provides additional insights into the summarization quality for this dataset.

These metrics were applied to both datasets, Qualco and FNS, to evaluate the performance of the summarization systems across different types of texts and usage contexts. By comparing the results from these two datasets, a more comprehensive understanding of the algorithms' effectiveness is achieved, highlighting potential areas for improvement.

Chapter 6

Results and Discussion

Chapter 6 focuses on analyzing the results of summaries generated by various models using the Qualco and FNS datasets. Initially, we perform a comparative analysis between the gold standard summaries created for a random sample of Qualco customers and the summaries produced by our models. Following this, we extend our evaluation to the entire Qualco dataset, comparing the model-generated summaries with the original customer comments to provide an alternative assessment of summary quality. Finally, we compare these results with the gold standard summaries from the FNS dataset to evaluate the models' performance across different datasets and assess their overall effectiveness.

6.1 Qualco SA Dataset

In this chapter, we assess the performance of the summarization models using data from Qualco. Our evaluation methodology involves a comparative analysis of the generated summaries against two distinct reference sets to provide a more nuanced assessment:

- **Comparison with Gold Summaries:** We assess the alignment between system-generated summaries and high-quality, human-created gold standard summaries. This step measures how well the model replicates expert-level summarization.
- **Comparison with Original Customer Comments:** We evaluate the effectiveness of the generated summaries by comparing them to the original,

unprocessed customer comments. This helps us gauge how well the summaries capture the key ideas and sentiments from the original feedback.

6.1.1 Subset with Gold Summaries

In table 6.1, we present the results of evaluating the gold standard summaries against the original customer comments. These gold summaries were created from 61 randomly selected samples, ensuring a diverse representation of customer interactions. This table shows how the quality of the gold summaries compares with the initial comments, giving us an estimation of the difficulty of the task. The scores reflect the degree of alignment between the gold summaries and the original customer feedback, as assessed by the ROUGE metrics.

Summaries	**Rouge-1**			**Rouge-2**			**Rouge-L**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Gold Summaries	0.99	0.69	0.80	0.91	0.56	0.68	0.99	0.69	0.80

Table 6.1: ROUGE scores for Gold Summaries against the Original Comments.

The results presented in Table 6.2 summarize the performance of all models—K-means, TF-IDF, and Random Sentences—tested with summaries of 5, 10, 15, and 20 sentences. As previously noted, the evaluation is based on 61 examples used to create the gold summaries. The data reveals a general trend: as the number of sentences in a summary increases, ROUGE scores tend to improve, especially in recall and F1 metrics. However, this increase in ROUGE scores may not necessarily indicate an improvement in summarization quality. It may instead highlight a limitation of the ROUGE metric, as longer summaries tend to achieve higher scores due to greater textual overlap with the reference document, irrespective of the actual accuracy or relevance of the generated summary.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	0.94	0.39	0.49	0.84	0.30	0.39	0.94	0.39	0.49
10 Sentences	0.97	0.63	0.72	0.91	0.54	0.63	0.97	0.63	0.72
15 Sentences	0.97	0.78	0.84	0.93	0.70	0.77	0.97	0.78	0.84
20 Sentences	0.97	0.86	0.90	0.95	0.80	0.85	0.97	0.86	0.90
TF-IDF									
5 Sentences	0.99	0.65	0.76	0.93	0.54	0.66	0.99	0.65	0.76
10 Sentences	0.99	0.82	0.89	0.91	0.71	0.78	0.99	0.82	0.89
15 Sentences	0.98	0.90	0.94	0.91	0.81	0.85	0.98	0.90	0.94
20 Sentences	0.98	0.95	0.96	0.92	0.87	0.89	0.98	0.95	0.96
Random Selection									
5 Sentences	1.00	0.45	0.57	0.88	0.36	0.46	1.00	0.45	0.57
10 Sentences	1.00	0.64	0.74	0.88	0.53	0.62	1.00	0.64	0.74
15 Sentences	1.00	0.77	0.84	0.91	0.67	0.74	1.00	0.77	0.84
20 Sentences	1.00	0.86	0.91	0.92	0.77	0.82	1.00	0.86	0.91

Table 6.2: ROUGE scores for Subset Dataset against the Original Comments.

The evaluation of summarization models reveals that random sentence selection consistently outperforms K-means in ROUGE-1 and ROUGE-L metrics across all summary lengths, highlighting its effectiveness despite its simplicity. In contrast, K-means demonstrates slightly superior performance in ROUGE-2 scores, indicating its capability to capture bi-gram structures better than random selection. TF-IDF shows the highest overall performance, particularly with longer summaries, achieving the best F1 scores across all ROUGE metrics.

These findings underscore the complexity of summarization as a task. While higher ROUGE scores may suggest improved performance, true quality in summarization transcends mere numerical evaluation. The differences in model effectiveness reveal the challenge of balancing precision, recall, and the subtleties of human language. This highlights the need for more comprehensive evaluation methods that better capture the nuances of summarization quality.

In the following table 6.3, we compare the generated abstracts with the golden abstracts to evaluate their performance.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	0.68	0.41	0.46	0.47	0.30	0.33	0.68	0.40	0.46
10 Sentences	0.68	0.64	0.63	0.52	0.51	0.49	0.67	0.63	0.62
15 Sentences	0.69	0.79	0.72	0.55	0.67	0.58	0.69	0.79	0.72
20 Sentences	0.69	0.86	0.75	0.55	0.76	0.62	0.69	0.86	0.75
TF-IDF									
5 Sentences	0.80	0.74	0.76	0.69	0.64	0.65	0.80	0.74	0.76
10 Sentences	0.76	0.91	0.82	0.64	0.81	0.70	0.76	0.91	0.82
15 Sentences	0.73	0.96	0.82	0.60	0.87	0.70	0.73	0.96	0.82
20 Sentences	0.70	0.97	0.81	0.57	0.89	0.69	0.70	0.97	0.81
Random Selection									
5 Sentences	0.70	0.45	0.52	0.52	0.36	0.39	0.69	0.45	0.52
10 Sentences	0.74	0.67	0.67	0.57	0.55	0.53	0.73	0.67	0.67
15 Sentences	0.70	0.76	0.71	0.54	0.64	0.57	0.69	0.76	0.71
20 Sentences	0.70	0.86	0.76	0.55	0.75	0.62	0.70	0.86	0.76

Table 6.3: ROUGE scores for Subset Dataset against the Gold Summaries.

In summary, all models exhibit improved ROUGE scores as the length of the summaries increases, indicating a general trend where longer summaries are better at capturing relevant information, as previously noted. Among the models tested, TF-IDF stands out as the most effective summarization technique, consistently outperforming both K-means and Random Sentences in alignment with gold summaries. This is particularly evident in the recall and F1 metrics, suggesting that TF-IDF not only produces summaries that align more closely with human-generated content but also retains key information from the original text.

Having previously established the methodology for calculating weights based on the timing of customer comments, we will now proceed to analyze the results with these weighted summaries 6.4, highlighting their advantages over the unweighted summaries discussed earlier.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	1.00	0.42	0.54	0.94	0.35	0.45	1.00	0.42	0.54
10 Sentences	1.00	0.62	0.72	0.97	0.56	0.66	1.00	0.62	0.72
15 Sentences	1.00	0.75	0.83	0.98	0.72	0.80	1.00	0.75	0.83
20 Sentences	1.00	0.84	0.89	0.99	0.81	0.87	1.00	0.84	0.89
TF-IDF									
5 Sentences	1.00	0.45	0.57	0.88	0.34	0.44	1.00	0.45	0.57
10 Sentences	1.00	0.64	0.74	0.85	0.51	0.60	1.00	0.64	0.74
15 Sentences	1.00	0.78	0.85	0.85	0.64	0.71	1.00	0.78	0.85
20 Sentences	1.00	0.85	0.90	0.85	0.70	0.75	1.00	0.85	0.90
Random Selection									
5 Sentences	1.00	0.20	0.30	0.79	0.14	0.21	1.00	0.20	0.30
10 Sentences	1.00	0.19	0.30	0.81	0.13	0.21	1.00	0.19	0.30
15 Sentences	1.00	0.20	0.30	0.80	0.14	0.21	1.00	0.20	0.30
20 Sentences	1.00	0.21	0.32	0.82	0.14	0.22	1.00	0.21	0.32

Table 6.4: ROUGE scores for Subset Dataset with Weights against the Original Comments.

It is observed that when weights are incorporated into the summaries, both K-means and TF-IDF demonstrate high accuracy and improved recall with longer summaries, indicating that a good percentage of relevant sentences are recovered and that the models are effective in capturing relevant content and aligning with the original annotations. This marks a significant improvement over their previous versions without weights. In contrast, the randomly selected summaries achieve high accuracy, but do not have the same level of recall and overall effectiveness, which highlights the limitations of this approach. Overall, the K-means and TF-IDF methods prove superior in generating comprehensive summaries that effectively represent customer feedback.

In Table 6.5, we will examine the performance of the models with weights in comparison to the gold summaries.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	0.69	0.43	0.50	0.51	0.33	0.37	0.69	0.43	0.49
10 Sentences	0.71	0.63	0.64	0.54	0.52	0.50	0.71	0.63	0.64
15 Sentences	0.70	0.76	0.71	0.55	0.66	0.58	0.70	0.75	0.71
20 Sentences	0.70	0.84	0.75	0.56	0.75	0.62	0.70	0.84	0.75
TF-IDF									
5 Sentences	0.71	0.47	0.53	0.54	0.36	0.41	0.71	0.47	0.53
10 Sentences	0.71	0.65	0.66	0.54	0.53	0.51	0.71	0.65	0.66
15 Sentences	0.72	0.79	0.74	0.55	0.67	0.59	0.72	0.79	0.74
20 Sentences	0.71	0.86	0.76	0.54	0.74	0.61	0.71	0.86	0.76
Random Selection									
5 Sentences	0.66	0.20	0.29	0.44	0.14	0.20	0.66	0.20	0.28
10 Sentences	0.70	0.20	0.29	0.46	0.14	0.19	0.70	0.20	0.28
15 Sentences	0.68	0.21	0.30	0.44	0.14	0.20	0.67	0.21	0.29
20 Sentences	0.71	0.22	0.32	0.48	0.16	0.22	0.71	0.22	0.31

Table 6.5: ROUGE scores for Subset Dataset with Weights against the Gold Summaries.

The evaluation shows that both K-means and TF-IDF generated summaries tend to improve precision and recall as the summary length increases, which is an indicator that a significant portion of relevant sentences is retrieved. In contrast, summaries from randomly selected sentences maintain high precision but demonstrate low recall, especially in shorter versions, highlighting their limited effectiveness. Overall, these findings emphasize the importance of summary length and selection method in text summarization, while cautioning against relying solely on ROUGE scores.

When comparing the results of Tables 6.2 and 6.4, it is evident that the introduction of weighting improves performance, particularly in terms of recall. Both tables show an increase in ROUGE scores with longer summaries, and the weighted summaries (Table: 6.4) consistently achieve higher F1 scores for both K-means and TF-IDF models. Additionally, the results in Tables 6.3 and 6.5 demonstrate that weighting enhances alignment with golden summaries, especially in recall and F1 scores. The TF-IDF model exhibits the most significant improvement, with weighted summaries outperforming unweighted ones across all metrics. K-means summaries also benefit from weighting, though to a lesser extent.

In conclusion, the introduction of weighting significantly enhances the performance of K-means and TF-IDF models when comparing summaries to the golden summaries, particularly in terms of recall and overall F1 scores. This improvement arises from the temporal relevance considered in the weighting process, which aligns better with the content generated by human experts. As a result, the models are more effective at capturing the essential information over time, leading to a more nuanced and accurate summarization.

6.1.2 Complete Dataset

In this section, we present the results of running the three selected models on the entire Qualco dataset. For this analysis, the previously held-out random samples, which were used for generating gold standard summaries, have been excluded to ensure that the evaluation is conducted on a distinct and comprehensive dataset. By applying the models to the full dataset, we aim to assess their performance across a broader range of customer interactions, providing a more generalized understanding of their effectiveness.

In Table 6.6, we present the results of evaluating the models without the influence of any weighting factors.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	1.00	0.80	0.86	0.94	0.73	0.78	1.00	0.80	0.86
10 Sentences	1.00	0.93	0.95	0.96	0.88	0.90	1.00	0.93	0.95
15 Sentences	1.00	0.97	0.98	0.97	0.93	0.94	1.00	0.97	0.98
20 Sentences	1.00	0.98	0.99	0.97	0.95	0.96	1.00	0.98	0.99
TF-IDF									
5 Sentences	1.00	0.89	0.93	0.95	0.83	0.87	1.00	0.89	0.93
10 Sentences	1.00	0.97	0.98	0.97	0.92	0.94	1.00	0.97	0.98
15 Sentences	1.00	0.98	0.99	0.97	0.95	0.96	1.00	0.98	0.99
20 Sentences	1.00	0.99	1.00	0.97	0.96	0.97	1.00	0.99	1.00
Random Selection									
5 Sentences	1.00	0.79	0.85	0.89	0.69	0.74	1.00	0.79	0.85
10 Sentences	1.00	0.91	0.94	0.93	0.85	0.88	1.00	0.91	0.94
15 Sentences	1.00	0.96	0.97	0.95	0.91	0.92	1.00	0.96	0.97
20 Sentences	1.00	0.98	0.99	0.96	0.94	0.95	1.00	0.98	0.99

Table 6.6: ROUGE scores for Complete Dataset against the Original Comments.

In summary, the analysis reveals that both K-Means and TF-IDF methods are effective for summarizing the dataset, with K-Means excelling in precision and TF-IDF demonstrating a balanced approach between accuracy and coverage. The results also highlight a potential limitation of the ROUGE metric, as longer summaries tend to achieve higher scores due to greater textual overlap with the reference documents, which may obscure the actual accuracy and relevance of the generated summaries.

K-Means improves its performance as the summary length increases, highlighting the significance of length in effective summarization. TF-IDF also demonstrates strong results, particularly with longer summaries, effectively capturing the structure and key details of the original content. In contrast, the random selection method, while precise, lacks coherence and often misses critical information.

In conclusion, both K-Means and TF-IDF are effective methods for summarizing the dataset, with K-Means being highly precise and TF-IDF offering a more balanced approach between precision and coverage. Random Selection, while simpler, is less consistent and may omit critical details, leading to lower overall quality in the summaries.

To extend our research, we applied weighting to the dataset and conducted ad-

ditional experiments. The results of these experiments are presented in Table 6.7.

Summaries	Rouge-1			Rouge-2			Rouge-L		
	P	R	F	P	R	F	P	R	F
K-Means									
5 Sentences	1.00	0.79	0.84	0.95	0.73	0.78	1.00	0.79	0.84
10 Sentences	1.00	0.91	0.94	0.97	0.87	0.90	1.00	0.91	0.94
15 Sentences	1.00	0.95	0.97	0.97	0.92	0.94	1.00	0.95	0.97
20 Sentences	1.00	0.97	0.98	0.97	0.94	0.95	1.00	0.97	0.98
TF-IDF									
5 Sentences	1.00	0.79	0.85	0.89	0.69	0.74	1.00	0.79	0.85
10 Sentences	1.00	0.92	0.94	0.84	0.76	0.79	1.00	0.92	0.94
15 Sentences	1.00	0.96	0.97	0.84	0.80	0.81	1.00	0.96	0.97
20 Sentences	1.00	0.98	0.98	0.84	0.82	0.82	1.00	0.98	0.98
Random Selection									
5 Sentences	1.00	0.78	0.84	0.95	0.73	0.78	1.00	0.78	0.84
10 Sentences	1.00	0.91	0.94	0.97	0.87	0.90	1.00	0.91	0.94
15 Sentences	1.00	0.95	0.97	0.97	0.92	0.94	1.00	0.95	0.97
20 Sentences	1.00	0.97	0.98	0.97	0.94	0.95	1.00	0.97	0.98

Table 6.7: ROUGE scores for Complete Dataset with Weights against the Original Comments.

The K-Means summarization method, enhanced by the incorporation of weights, consistently outperforms other techniques in generating high-quality summaries, excelling in both content inclusion and completeness. Its ability to capture essential details improves significantly with longer summaries, making it a robust choice for effective summarization. Similarly, the TF-IDF method demonstrates strong precision but tends to have lower recall compared to K-Means, indicating it may miss some critical content. While the Random Selection method achieves similar precision, it lacks the coherence and reliability necessary for comprehensive summarization, resulting in less effective content capture.

In summary, the evaluation of the K-Means, TF-IDF, and Random Selection summarization methods across various ROUGE metrics highlights distinct strengths. K-Means excels in both precision and completeness, making it the most robust method for generating comprehensive summaries, particularly for longer texts. TF-IDF shows strong precision but lower recall, indicating it captures less content overall

compared to K-Means. Random Selection, while precise, demonstrates more variability and inconsistency in summary quality. Overall, K-Means is the preferred approach, though TF-IDF and Random Selection may still be useful depending on specific needs for precision versus completeness.

6.2 FNS Dataset

This section presents the results of our summarization model's performance across financial documents in Greek, English, and Spanish, utilizing the dataset from the Financial Narrative Summarisation (FNS) 2023 task [24]. The goal is to evaluate the model's effectiveness in generating 1000-word summaries and compare these summaries with the gold-standard reference summaries provided by the FNS 2023. We focus on the model's ability to handle the unique challenges posed by different languages and document structures, assessing its overall accuracy, coherence, and adaptability.

In the following, we present the evaluation results for the three datasets provided by the FNS organizers and compare our model's performance with other systems submitted in the context of the FNS shared task. More information regarding the participating systems can be found in [24].

6.2.1 Greek Financial Documents

Our initial experiments were conducted on Greek financial reports due to their complex and less standardized nature. The K-means clustering algorithm was employed to generate 1000-word summaries from these reports. The results indicated that the model was able to identify and cluster key thematic elements within the documents effectively.

The model captured the main financial metrics and narrative themes. However, due to the unstructured nature of the Greek reports, the summaries sometimes included redundant information, reflecting the variability in the source material. The summaries generally maintained a logical flow, though certain sections lacked the fluidity found in more structured documents. This was particularly evident in areas where financial data was integrated with narrative content, leading to occasional

disjointedness.

Evaluation using ROUGE metrics [31]showed that the K-Means model performed with mixed results, as seen in Table 6.8,which also presents the performance of the systems that participated in the FNS task.

Team_System	Rouge-1			Rouge-2			Rouge-L			Rouge-SU4		
	P	R	F	P	R	F	P	R	F	P	R	F
FNS 2023 Results - Greek Dataset												
SCE bertplm	0.44	0.25	0.32	0.21	0.10	0.13	0.36	0.21	0.26	0.25	0.13	0.17
SCE plm	0.44	0.25	0.32	0.21	0.10	0.13	0.36	0.21	0.26	0.27	0.13	0.17
Rocky T5	0.48	0.23	0.31	0.19	0.10	0.13	0.35	0.20	0.25	0.23	0.12	0.16
SSC AI RG DiMSum	0.44	0.22	0.29	0.25	0.08	0.12	0.33	0.16	0.20	0.26	0.12	0.16
MBertExtractive baseline	0.26	0.19	0.22	0.08	0.04	0.05	0.10	0.11	0.10	0.12	0.09	0.10
K-Means												
K-Means	0.43	0.26	0.32	0.19	0.07	0.10	0.35	0.19	0.25	0.25	0.12	0.16

Table 6.8: ROUGE scores for FNS Dataset comparing K-Means summaries against Gold Summaries in the Greek Dataset.

The K-Means method successfully identifies key terms from the Greek financial reports but struggles to capture the full scope of the content, which impacts the coherence and structure of the summaries. There is a noticeable pattern where the model can cluster central themes but falls short in conveying the complete narrative of the original text. Additionally, more complex metrics that account for word pairs reveal similar shortcomings. While K-Means is useful for clustering important ideas, it does not maintain the flow or overall structure of the document. In contrast, models like "SCE bertplm" and "Rocky T5" provide more balanced summaries, capturing both essential content and greater coherence.

This suggests that improvements to K-Means, or the integration of complementary techniques, may be needed to achieve a better balance between precision and coverage, ensuring that both relevance and structure are effectively represented in the generated summaries.

6.2.2 English Financial Documents

We subsequently applied the K-means summary model to English financial reports, which are usually more structured than their Greek counterparts. The performance of the model on these documents was significantly better as the summaries effectively captured the key financial narratives and key data points, aligning closely with the gold standard summaries from the FNS 2023 document. Due to the structured nature of the English reports, the resulting summaries were more coherent and logically organized, with fewer instances of redundancy.

Based on the results presented in Table 6.9, we can draw several conclusions regarding the performance of the K-Means summarization model in comparison to other models on the English dataset from the FNS 2023 task.

Team_System	**Rouge-1**			**Rouge-2**			**Rouge-L**			**Rouge-SU4**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
FNS 2023 Results - English Dataset												
SCE bertplm	0.43	0.44	0.42	0.26	0.28	0.26	0.38	0.45	0.41	0.30	0.33	0.30
SCE plm	0.44	0.46	0.43	0.28	0.30	0.27	0.39	0.46	0.41	0.33	0.35	0.33
Rocky T5	0.27	0.25	0.24	0.11	0.15	0.12	0.50	0.19	0.26	0.12	0.19	0.14
SSC AI RG DiMSum	0.48	0.53	0.48	0.27	0.44	0.32	0.46	0.51	0.47	0.24	0.48	0.31
SSC AI RG DiMSum1	0.48	0.37	0.38	0.23	0.24	0.22	0.44	0.36	0.37	0.23	0.29	0.24
SSC AI RG DiMSum2	0.38	0.48	0.40	0.21	0.33	0.24	0.36	0.44	0.38	0.24	0.38	0.28
MBertExtractive baseline	0.31	0.27	0.28	0.11	0.14	0.11	0.23	0.24	0.23	0.12	0.19	0.14
K-Means												
K-Means	0.25	0.27	0.25	0.09	0.10	0.09	0.23	0.22	0.21	0.13	0.16	0.14

Table 6.9: ROUGE scores for FNS Dataset comparing K-Means summaries against Gold Summaries in the English Dataset.

The K-Means method applied to the English dataset shows mixed results. While it manages to capture some key terms from the reference summaries, it struggles to encompass the full range of relevant content, resulting in summaries that lack depth and coherence. The model performs particularly weakly when it comes to capturing more complex content relationships, which is reflected in its difficulty

with metrics that account for bigram overlap and sentence structure. Although useful for clustering important themes, K-Means fails to maintain the structural integrity and flow of the original text, making the summaries less comprehensive. In contrast, models like "SCE bertplm" and "SSC AI RG DiMSum" offer better balance between content relevance and coherence, producing summaries that are more reflective of the source material.

Overall, the performance of K-Means on the English dataset highlights the need for further refinement or the use of additional methods to improve the completeness and clarity of the generated summaries. Overall, the results obtained from the K-Means method applied to the English dataset indicate a clear opportunity for enhancement.

6.2.3 Spanish Financial Documents

Finally, we extended our analysis to Spanish financial reports. These reports often feature a narrative style that is more descriptive, with a tendency to contextualize financial data within broader economic and business discussions. The K-means model performed well in identifying key themes, although the descriptive nature of the text sometimes led to summaries that were less concise compared to the reference summaries. In general, the summaries were coherent but sometimes suffered from the over-inclusion of contextual information which, although relevant, distracted from the focus on the key financial metrics.

Given the results presented in Table 6.10, we can derive several insights into the performance of different summarization models on the Spanish dataset.

Team_System	Rouge-1			Rouge-2			Rouge-L			Rouge-SU4		
	P	R	F	P	R	F	P	R	F	P	R	F
FNS 2023 Results - Spanish Dataset												
SCE bertplm	0.34	0.51	0.36	0.12	0.20	0.13	0.20	0.27	0.21	0.17	0.27	0.18
SCE plm	0.40	0.46	0.41	0.14	0.15	0.14	0.25	0.27	0.25	0.20	0.21	0.20
Rocky T5	0.45	0.36	0.39	0.09	0.08	0.08	0.15	0.16	0.15	0.16	0.14	0.14
SSC AI RG DiMSum	0.37	0.45	0.40	0.11	0.13	0.11	0.16	0.17	0.16	0.16	0.20	0.17
MBertExtractive baseline	0.38	0.36	0.36	0.08	0.08	0.08	0.14	0.14	0.14	0.15	0.15	0.15
K-Means												
K-Means	0.37	0.44	0.39	0.09	0.11	0.10	0.15	0.17	0.16	0.15	0.19	0.16

Table 6.10: ROUGE scores for FNS Dataset comparing K-Means summaries against Gold Summaries in the Spanish Dataset.

The evaluation of the K-Means model on the Spanish dataset shows mixed results across ROUGE metrics. While it achieves moderate ROUGE-1 Precision and Recall scores, its performance in ROUGE-2 and ROUGE-L reveals challenges in capturing detailed content and maintaining coherence. Compared to other models like "SCE bertplm" and "SCE plm," which demonstrate stronger performance, K-Means highlights the need for further refinement or complementary techniques to enhance the quality and comprehensiveness of the generated summaries.

Summarizing the results from the three datasets (Greek, English, and Spanish) of the FNS 2023 competition, it is observed that the K-Means model, which is the focus of this work, exhibited consistent performance across all datasets, showing a tendency towards moderate precision and recall. Although it did not achieve the highest performance compared to more advanced models, the results suggest that K-Means can serve as a reliable baseline for text summarization, particularly in multilingual datasets. The fact that K-Means remained competitive across all languages indicates its potential as a simple yet powerful tool in financial text processing, while also highlighting areas for improvement to achieve more accurate and comprehensive summaries.

Chapter 7

Conclusions

In this work, we studied extractive summarization techniques focusing on applications in the financial domain. Specifically, we utilized a conversation-based dataset from Qualco and analyzed various summarization techniques using K-Means clustering, TF-IDF, and random sentence selection. Our primary objective was to develop a model capable of grouping similar sentences, as the Qualco dataset contained a significant amount of unstructured and repetitive content. By applying this approach to financial documents in multiple languages—specifically English, Greek, and Spanish—we aimed to evaluate the K-Means model's effectiveness in summarizing complex financial texts and to ascertain whether it could successfully capture and convey essential information.

Our contributions include a thorough analysis of the strengths and weaknesses of various summarization techniques, particularly highlighting the performance of utilizing K-Means in comparison to other approaches, such as TF-IDF and random sentence selection. Our findings indicate that while K-Means serves as a robust baseline, its effectiveness varies significantly across languages due to differences in document structure and narrative style. Specifically, the model achieved relatively good performance with Greek and Spanish texts but struggled with English documents, possibly due to differences in text complexity and the lack of a standardized structure in the dataset. These results highlight the importance of tailoring summarization techniques to the linguistic and structural characteristics of the target language.

Furthermore, our analysis revealed a trend wherein longer summaries achieved higher ROUGE scores due to greater textual overlap with reference documents. This observation may highlight a limitation of the ROUGE metric itself, as it does not always reflect the actual accuracy or relevance of the generated summaries. While simpler methods like TF-IDF balance precision and recall, they often miss critical details. Moreover, it became evident how the effectiveness of the model can be heavily influenced by the nature of the input data, as well as by the specific characteristics of different languages. This suggests the need for ongoing refinement of summarization techniques to enhance their performance across diverse financial datasets.

Looking ahead, future research could investigate hybrid approaches that integrate the strengths of K-Means with more advanced models, such as transformer-based architectures, to enhance the summarization of unstructured and multilingual financial data. Exploring these methodologies has the potential to yield more accurate and informative summaries, ultimately improving decision-making processes within the Fintech sector.

References

[1] K. B. Prasasthy. Brief history of text summarization. *Medium*, 2021.

[2] H. P. Luhn. The automatic creation of literature abstracts. *IBM JOURNAL*, 1958.

[3] Prateek Joshi. An introduction to text summarization using the textrank algorithm (with python implementation). *Analytics Vidhya*, 2023.

[4] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[5] Mitesh Dewda. Abstractive text summarization. *Medium*, 2022.

[6] L. Li, C. Forascu, and G. El-Haj, M. andGiannakopoulos. Multi-document multilingual summarization corpus preparation, part 1:arabic, english, greek, chinese, romanian. *MultiLing 2013 Workshop, held within the ACL 2013 Conference*, 2013.

[7] Vishal Gupta and Gurpreet Lehal. A survey of text summarization extractive techniques. *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3*, 2010.

[8] E. Kantzola. Extractive text summarization of greek news articles based on sentence-clusters. *Uppsala University*, 2020.

[9] G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 2008.

[10] G. Giannakopoulos, V. Karkaletsis, and G. Vouros. Automatic summarization and background knowledge: Past, present and vision technical report (submission no. demo-2006-2). 2006.

[11] M. Elhadad. Using argumentation to control lexical choice: A functional unification-based approach. *Columbia University*, 1992.

[12] S. Gholamrezazadeh, M.A. Salehi, and B. Gholamzadeh. A comprehensive survey on text summarization systems. *IEEE*, 2009.

[13] R. Khan, Y. Qian, and S. Naeem. Extractive based text summarization using k-means and tf-id. *I.J. Information Engineering and Electronic Business*, 2019.

[14] J. Xu and G. Durrett. Neural extractive text summarization with syntactic compression. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[15] W.Z. AL-Dyani, F.K. Ahmad, and S.S. Kamaruddin. A survey on event detection models for text data streams. *Journal of Computer Science*, 2020.

[16] J. Neto, A. Freitas, and C.A.A. Kaestner. Automatic text summarization using a machine learning approach. *Lecture Notes in computer science, Springer Berlin / Heidelberg*, 2002.

[17] B. Kumar, U.K. Tiwari, and D.C. Dobhal. Machine learning based approach for user story clustering in agile engineering. *https://doi.org/10.1007/s42979-023-02212-2*, 2023.

[18] D.F. Coimbra. Hybrid extractive/abstractive summarization using pre-trained sequence-to-sequence models. *https://doi.org/10.1016/j.eswa.2021.116292*, 2020.

[19] P.J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[20] C. Y. Lin. Rouge: A package for automatic evaluation of summaries. *University of Southern California*, 2004.

[21] K. Papineni, S. Roukos, T. Ward, and W. ZHU. Bleu: A method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2001.

[22] J.N Madhuri and R. Ganesh Kumar. Extractive text summarization using sentence ranking. *IEEE*, 2019.

[23] J. Steinberger and K. Jezek. Evaluation measures for text summarization. *University of West Bohemia in Pilsen*, 2007.

[24] E. Zavitsanos, A. Kosmopoulos, G. Giannakopoulos, M. Litvak, B. Carbajo-Coronado, A. Moreno-Sandoval, and M. El-Ha. The financial narrative summarisation shared task (fns 2023). *2023 IEEE International Conference on Big Data (BigData)*, 2023.

[25] S. Everette and Jr. Gardner. Exponential smoothing: The state of the art. *https://doi.org/10.1002/for.3980040103*, 1985.

[26] A. Abrami, A. Y. Aravkin, and Y. Kim. Time series using exponential smoothing cells. *https://doi.org/10.48550/arXiv.1706.02829*, 2017.

[27] R. Weron, K. Weron, and A. Weron. A conditionally exponential decay approach to scaling in finance. *https://doi.org/10.1016/S0378-4371(98)00547-0*, 1999.

[28] E. Lloret, L. Plaza, and A. Aker. The challenging task of summary evaluation: An overview. *https://doi.org/10.1007/s10579-017-9399-2*, 2017.

[29] A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold standard. *https://doi.org/10.1162/COLI*$_{a0}$0123, 2013.

[30] F. Falcão. Metrics for evaluating summarization of texts performed by transformers: how to evaluate the quality of summaries. *Medium*, 2023.

[31] P. Fung and G. Ngai. One story, one flow: Hidden markov story models for multilingual multidocument summarisation. *ACM Transactions on Speech and Language Processing (TSLP), 3(2):1–16*, 2006.