

**EXTRACTIVE TEXTRANK-BASED NLP NEWS SUMMARIZATION FOR
MULTIPLE DOMAINS**

BY

**MD WAZIH ULLAH MUSTOFA
ID: 192-15-13149**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD Ashrafal Islam Talukder
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled “**Extractive TextRank-Based NLP News Summarization For Multiple Domains**” submitted by **Student ID: 192-15-13149**, to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26 January, 2024.

BOARD OF EXAMINERS

Chairman

Dr. Md. Zahid Hasan (ZH)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher (THT)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Abbas Ali-Khan (AAK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

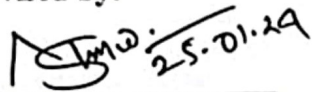
External Examiner

Dr. Mohammed Nasir Uddin (DNU)
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

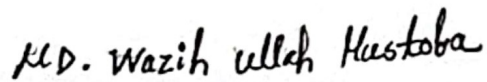
I hereby declare that this project has been done by me under the supervision of **Md Ashraful Islam Talukder, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Md Ashraful Islam Talukder
Lecturer
Department of CSE
Daffodil International University

Submitted by:



MD Wazih Ullah Mustofa
ID: -192-15-13149
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish my profound my indebtedness to **MD Ashraful Islam Talukder, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of Data Mining, Machine Learning (ML), Deep learning, Natural language processing (NLP) to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheikh Rashed Haider Noori, Professor, and Head**, Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

This paper provides a thorough analysis of extractive summarization, or the use of Natural Language Processing (NLP) techniques to summarize news articles. Approximately two thousand articles covering a wide range of topics, including business, entertainment, politics, sports, and technology, were gathered from different online platforms, including the well-known "Prothom Alo" newspaper. My method included a thorough preprocessing step that included punctuation and special character removal, as well as spell correction with TextBlob. The primary focus of my study is the implementation of the TextRank algorithm, which was modified from the PageRank algorithm to handle natural language text. Using this technique, text was represented as a graph, with edges denoting the cosine similarity between sentences and vertices representing the sentences themselves. I described my process for vectorizing sentences and creating a similarity matrix by figuring out the cosine similarity between each pair. The paper explores the algorithmic nuances of using a customized sentence similarity function to rank sentences according to their relevance and importance. I then conducted a comparative analysis of the summaries generated against the original texts, calculating similarity scores to evaluate the efficacy of my summarization process. The study aims to highlight the effectiveness of extractive summarization in processing large volumes of news data, offering insights into the potential of NLP in media analytics. By comparing the actual summaries and those generated through my method, I draw conclusions about the precision and utility of extractive summarization in the context of diverse news content. This research contributes to the field by demonstrating a practical application of NLP in the efficient processing and summarization of large-scale news data.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	1-9
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.3.1 Advancement of NLP Techniques	3
1.3.2 Improved Information Retrieval	3
1.3.3 Relevance to News Aggregation	3
1.3.4 Addressing the Big Data Challenge	4
1.3.5 Adaptation to Evolving Writing Styles	4
1.4 Research Questions	4
1.5 Expected Output	5
1.5.1 Development of an Efficient Extractive Summarization Model	5
1.5.2 Evaluation of TextRank Algorithm Effectiveness	6
1.5.3 Insights into Preprocessing Impact	6
1.5.4 Comparative Analysis	6

1.5.5 Category-Wise Performance Analysis	6
1.5.6 Recommendations for Future Research	6
1.5.7 Contribution to NLP and Text Summarization	6
1.6 Report Layout	7
CHAPTER 2: BACKGROUND	10-20
2.1 Preliminaries/Terminologies	10
2.1.1 Extractive Text Summarization	10
2.1.2 TextRank Algorithm	10
2.1.3 Preprocessing Techniques	11
2.1.4 Similarity Scores	11
2.1.5 Category-Wise Performance	11
2.1.6 Benchmark Summarizer (Summy)	12
2.2 Related Works	12
2.3 Comparative Analysis and Summary	15
2.4 Scope of the Problem	17
2.5 Challenges	19
CHAPTER 3: RESEARCH METHODOLOGY	21-29
3.1 Research Subject and Instrumentation	21
3.2 Data Collection Procedure/Dataset Utilized	23
3.3 Statistical Analysis	25
3.4 Proposed Methodology/Applied Mechanism	26
3.5 Implementation Requirements	27
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	30-37

4.1 Experimental Setup	30
4.2 Experimental Results & Analysis	31
4.2.1 Average Similarity Scores for Preprocessing Levels	31
4.2.2 Average Similarity Scores for Preprocessing Levels (Similarity > 0.5)	32
4.2.3 Average Similarity Scores for Category-Wise Analysis (Spell-Corrected News)	34
4.2.4 Average Similarity Scores for Category-Wise Analysis (Fully Preprocessed News)	35
4.3 Discussion	36
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	38-43
5.1 Impact on Society	38
5.2 Impact on Environment	39
5.3 Ethical Aspects	40
5.4 Sustainability Plan	42
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	44-48
6.1 Summary of the Study	44
6.2 Conclusions	45
6.3 Implication for Further Study	46
APPENDIX	49-50
REFERENCES	51-53
PLAGIARISM REPORT	54

LIST OF FIGURES:

FIGURES	PAGE NO
Figure 2.1.2.1: TextRank Algorithm	10
Figure 2.1.3.1: Dataset Preprocessing	11
Figure 2.1.6.1: Summy Summarizer	12
Figure 2.3.1: Comparative Analysis	16
Figure 3.1.1: Process of Research Methodology	21
Figure 3.1.2: Cosine Similarity Matrix	23
Figure 3.2.1: Data collection	24
Figure 4.2.1.1: Visual Representation of similarity scores for different preprocessing levels.	32
Figure 4.2.2.1: Visual Representation of average similarity scores	33
Figure 4.2.3.1: Visual representation of average similarity score for spell-corrected news	34
Figure. 4.2.4.1: Visual representation of average similarity scores for Fully Processed News	35

LIST OF TABLES:

TABLES	PAGE NO
Table 4.2.1: A similarity scores for different preprocessing levels.	32
Table 4.2.2 A closer look at average similarity scores.	33
Table 4.2.3 Average similarity scores for spell-corrected news.	34
Table 4.2.4 Average Similarity Scores for Category-Wise Analysis (Fully Preprocessed News)	35

CHAPTER 1

Introduction

1.1 Introduction

The digital age has ushered in an era of unprecedented information abundance, primarily in the form of online news. This proliferation of content, while advantageous, presents a significant challenge in terms of information management and comprehension. Extractive summarization, an essential branch of Natural Language Processing (NLP), emerges as a vital tool in this context. It involves selecting key sentences or fragments from a text to create a condensed version that effectively conveys the main ideas. This technique is crucial in fields such as news aggregation, academic research, and business intelligence, where quick and accurate understanding of vast textual information is essential.

The impetus for this research is rooted in the ever-increasing volume of English-language news content and the necessity for sophisticated summarization technologies. Despite English being extensively studied in the realm of NLP, the continuous evolution in the style and diversity of news writing poses unique challenges. Our study focuses on addressing these challenges by developing an efficient extractive summarization model tailored for English news articles. The research employs a comprehensive dataset drawn from various online news platforms, encompassing a range of topics, including business, entertainment, politics, sports, and technology.

The primary objectives of this study are threefold:

To implement and evaluate the effectiveness of the TextRank algorithm and Cosine Similarity measures in the context of summarizing English news articles.

To compare the performance of our model against manual summarization and a benchmark built-in summarizer, Summy, in terms of accuracy and coherence.

To conduct a detailed analysis of the model's performance across different news categories, offering insights into its adaptability and robustness.

This paper is structured to provide a comprehensive narrative of the research undertaken. Chapter 2 delves into the background, encompassing a literature review and the theoretical underpinnings of the methodologies used. Chapter 3 outlines the research methodology, detailing the processes of data collection, preprocessing, and the approach to extractive summarization. Chapter 4 presents the experimental results, followed by a discussion of

their implications in the broader context of NLP and news summarization. Finally, the paper concludes with a summary of the key findings and perspectives on potential future research directions in this evolving field.

1.2 Motivation

The motivation behind this research is rooted in the transformative impact of the digital era on the way information is disseminated and consumed. With the internet and digital platforms becoming the primary sources of news and information, there has been an exponential increase in the volume of textual content available to individuals and organizations. While this abundance of information is a testament to the power of the digital age, it also presents a significant challenge - the need for effective information management.

In this information-rich environment, the ability to distill the essence of lengthy texts quickly and accurately is more critical than ever. Whether it's journalists sifting through a vast array of news articles, researchers conducting literature reviews, or business professionals seeking insights from textual data, the demand for efficient text summarization techniques is pronounced.

One of the most common scenarios where text summarization is indispensable is in the realm of news aggregation. Online news platforms continuously generate articles across a multitude of topics, and readers often need to grasp the key points without investing substantial time in reading each article in its entirety. In academia, researchers encounter a similar challenge when reviewing an extensive body of literature. Extracting essential information from research papers and articles can be time-consuming and labor-intensive. Furthermore, in the business world, the ability to distill actionable insights from reports, market analyses, and business news articles is integral to making informed decisions. The sheer volume of textual data available in this context necessitates automated text summarization methods that can extract the most pertinent information efficiently.

The motivation for this research extends to the exploration of how advancements in Natural Language Processing (NLP) can address these challenges. By focusing on the development and evaluation of an extractive summarization model tailored for English news articles, this research seeks to contribute to the broader field of NLP. It aims to enhance the capabilities of machines in understanding and summarizing human language effectively.

Ultimately, the motivation behind this study lies in the potential to empower individuals, organizations, and researchers with tools that can navigate the sea of textual data, distill the critical information, and enable more informed decision-making. By developing and evaluating an efficient extractive summarization model, this research endeavors to bridge the gap between the ever-expanding world of digital information and the need for concise, meaningful insights.

1.3 Rationale of the Study

The rationale for undertaking this research is grounded in the fundamental challenges posed by the information age, where the exponential growth of textual data has created a pressing need for effective summarization techniques. In this section, we outline the key rationales that drive this study

1.3.1 Advancement of NLP Techniques

One of the primary rationales for this research is the aspiration to contribute to the advancement of Natural Language Processing (NLP) techniques, particularly in the domain of text summarization. As NLP continues to evolve, the development of more sophisticated and accurate summarization models holds immense value. This research seeks to explore innovative methodologies and strategies that can enhance the capabilities of NLP in understanding and summarizing complex human language.

1.3.2 Improved Information Retrieval

The exponential growth of digital information has created a significant challenge in terms of information retrieval and management. Efficient and accurate text summarization methods are pivotal in addressing this challenge. By condensing lengthy texts into concise summaries, individuals, researchers, and organizations can access and comprehend vast volumes of information more effectively. This research aims to contribute to the development of tools and techniques that improve information retrieval and knowledge management.

1.3.3 Relevance to News Aggregation:

With the proliferation of online news sources, news aggregation platforms have become indispensable for keeping up with current events. These platforms rely on summarization algorithms to provide users with quick and informative summaries of news articles. The effectiveness of these algorithms directly impacts the user experience. Therefore, this

research holds relevance to news aggregation platforms, offering insights into how extractive summarization can enhance the presentation of news content.

1.3.4 Addressing the Big Data Challenge:

In an era characterized by the explosion of big data, the ability to process and distill meaningful insights from vast textual datasets is crucial. Extractive summarization techniques can help mitigate information overload by identifying and presenting the most relevant and informative portions of texts. This research addresses the pressing need for tools that can facilitate the management of big data by efficiently summarizing textual content.

1.3.5 Adaptation to Evolving Writing Styles:

The world of news reporting and content creation is dynamic, with writing styles and formats constantly evolving. This research recognizes the importance of developing summarization techniques that can adapt to these changes. By exploring the effectiveness of the TextRank algorithm and preprocessing techniques, this study aims to assess the model's ability to handle diverse writing styles and evolving patterns in news articles.

In summary, the rationale for this research is multifaceted, encompassing contributions to NLP, improvements in information retrieval, relevance to news aggregation, addressing the challenges of big data, and adaptability to evolving writing styles. These rationales collectively underscore the significance and relevance of this study in the context of the digital age and the ever-expanding realm of textual information.

1.4 Research Questions:

The research questions serve as the guiding compass for this study, directing the focus and investigation into the field of extractive text summarization. In this section, we delineate the primary research questions that shape the trajectory of this research:

- **How effective is the TextRank algorithm in the context of extractive summarization for English news articles?**

This foundational research question delves into the core of the study. It seeks to assess the efficacy of the TextRank algorithm, a prominent graph-based ranking algorithm inspired by Google's PageRank, in the specific context of summarizing English news articles. The investigation explores the algorithm's ability to identify and rank key sentences or fragments within news articles accurately.

- **What impact do different preprocessing techniques have on the performance of extractive summarization models for news articles?**

This research question shifts the focus to preprocessing techniques, a critical aspect of text summarization. It seeks to understand how various preprocessing methods, including spell correction, removal of special characters, normalization, and sentence segmentation, influence the overall performance of extractive summarization models when applied to news articles. The goal is to uncover the significance of preprocessing in enhancing the quality of summarization.

- **How does the performance of the proposed extractive summarization model compare to manually generated summaries and a benchmark summarizer like Summy in terms of accuracy and coherence?**

This comparative research question extends the evaluation beyond the algorithm's technical effectiveness. It assesses the model's summarization output by comparing it to two benchmarks: manually generated summaries crafted by experts and summaries produced by Summy, a built-in summarization tool. The evaluation criteria include accuracy, coherence, completeness, and conciseness. This question provides insights into the model's ability to mimic human summarization skills and its performance relative to established summarization methods.

These research questions collectively form the framework for this study, guiding the research methodology, experimentation, and analysis. They represent the critical inquiries that drive the exploration of extractive text summarization within the domain of English news articles.

1.5 Expected Output

The expected output of this research encompasses several key deliverables and anticipated outcomes that align with the research objectives and questions:

1.5.1 Development of an Efficient Extractive Summarization Model:

One of the primary expected outputs of this research is the development of an efficient extractive summarization model tailored for English news articles. This model is anticipated to incorporate the TextRank algorithm and preprocessing techniques to achieve accurate and concise summaries.

1.5.2 Evaluation of TextRank Algorithm Effectiveness:

The research aims to provide insights into the effectiveness of the TextRank algorithm in the context of extractive summarization. The expected output includes a comprehensive evaluation of the algorithm's performance, including its ability to identify and rank key sentences or fragments within news articles.

1.5.3 Insights into Preprocessing Impact:

The research anticipates providing valuable insights into the impact of different preprocessing techniques on extractive summarization model performance. This includes the assessment of how spell correction, special character removal, text normalization, and sentence segmentation influence the quality of summarization outputs.

1.5.4 Comparative Analysis:

The comparative analysis between the proposed extractive summarization model, manually generated summaries, and Summy-generated summaries is a significant expected output. It aims to determine how the model's accuracy and coherence measure up against human-generated summaries and an established summarization tool.

1.5.5 Category-Wise Performance Analysis:

The research is expected to yield category-wise performance analyses, providing insights into how the summarization model performs across different news categories, including business, entertainment, politics, sports, and technology. This analysis aims to showcase the model's adaptability and robustness.

1.5.6 Recommendations for Future Research:

The study is likely to generate recommendations for future research directions in the field of extractive summarization. These recommendations may include potential enhancements to the summarization model, the exploration of additional algorithms, and novel approaches to handling evolving writing styles.

1.5.7 Contribution to NLP and Text Summarization:

Ultimately, the expected output of this research extends to its contribution to the broader field of Natural Language Processing (NLP) and text summarization. It aims to advance the state-of-the-art in NLP by introducing a model tailored for news articles and by shedding light on the nuances of extractive summarization in this domain.

These expected outputs collectively represent the research's intended contributions to knowledge, technology, and the understanding of text summarization in the context of English news articles. They serve as the tangible outcomes that align with the study's objectives and are poised to impact various sectors, including journalism, academia, and information retrieval.

1.6 Report Layout

The layout of this report is designed to provide a structured and comprehensive overview of the research conducted on extractive summarization for English news articles. The following sections outline the report's layout and the content covered in each chapter:

Chapter 2: Background

2.1 Preliminaries/Terminologies: This section introduces essential concepts and terminologies relevant to the study, ensuring a clear understanding of the subsequent chapters.

2.2 Related Works: This section reviews existing literature and studies in the domain of extractive text summarization, providing insights into the state of the field.

2.3 Comparative Analysis and Summary: This part summarizes and compares the methodologies and approaches used in prior research, highlighting key findings and gaps in knowledge.

2.4 Scope of the Problem: Here, the scope and limitations of the research are defined, outlining the specific focus areas and objectives.

2.5 Challenges: This section discusses the challenges and complexities associated with extractive summarization for news articles, setting the context for the research.

Chapter 3: Research Methodology

3.1 Research Subject and Instrumentation: This section describes the research subject, including the TextRank algorithm and preprocessing techniques. It also outlines the research instrumentation, including data sources and tools.

3.2 Data Collection Procedure/Dataset Utilized: Details about data collection procedures and the dataset used in the research are presented here.

3.3 Statistical Analysis: This part discusses the statistical methods and analyses employed in the research, including similarity scores and category-wise performance evaluations.

3.4 Proposed Methodology/Applied Mechanism: The methodology applied in the research, encompassing the use of the TextRank algorithm and preprocessing techniques, is elaborated upon in this section.

3.5 Implementation Requirements: This section outlines the requirements for implementing the research, including hardware, software, and data sources.

Chapter 4: Experimental Results and Discussion

4.1 Experimental Setup: Details of the experimental setup, including data preprocessing, model development, and evaluation parameters, are presented here.

4.2 Experimental Results & Analysis: This section provides an in-depth analysis of the experimental results, including average similarity scores, category-wise performance, and comparisons with benchmark summarizers.

4.3 Discussion: The discussion section interprets the findings, analyzes the implications, and offers insights into the effectiveness of the extractive summarization model.

Chapter 5: Impact on Society, Environment and Sustainability

5.1 Impact on Society: This section explores the societal implications of the research, including its potential benefits for individuals and organizations in information retrieval and decision-making.

5.2 Impact on Environment: Environmental considerations related to the research, including resource utilization and sustainability, are discussed here.

5.3 Ethical Aspects: Ethical considerations associated with data handling, privacy, and informed consent are addressed in this part.

5.4 Sustainability Plan: This section outlines the sustainability plan for the research, emphasizing long-term impact and relevance.

Chapter 6: Summary, Conclusion, Recommendation, and Implication for Future Research

6.1 Summary of the Study: A concise summary of the research's key findings and contributions is presented here.

6.2 Conclusions: This section offers conclusive remarks, drawing from the research's outcomes and their significance.

6.3 Implication for Further Study: Recommendations for future research directions and potential enhancements to the summarization model are discussed in this part.

Each chapter is structured to provide a logical flow of information, ensuring that readers can navigate through the report seamlessly and gain a comprehensive understanding of the research, its methodologies, and its outcomes.

CHAPTER 2: Background

2.1 Preliminaries/Terminologies

This section of the report serves as a foundation for understanding the key concepts and terminologies essential to the field of extractive text summarization. These preliminary explanations ensure clarity and facilitate comprehension throughout the subsequent chapters.

2.1.1 Extractive Text Summarization:

Extractive text summarization is a technique within natural language processing (NLP) that involves the automatic selection of sentences or phrases from a source text to create a condensed and coherent summary. The selected content is extracted directly from the original text, and the goal is to capture the most important information while maintaining the original context.

2.1.2 TextRank Algorithm:

The TextRank algorithm is a prominent graph-based ranking algorithm used for extractive summarization. Inspired by Google's PageRank algorithm, TextRank assesses the importance of sentences within a text based on their relationships with other sentences. Important sentences are those that are connected to other important sentences, reflecting their significance in conveying the main ideas.

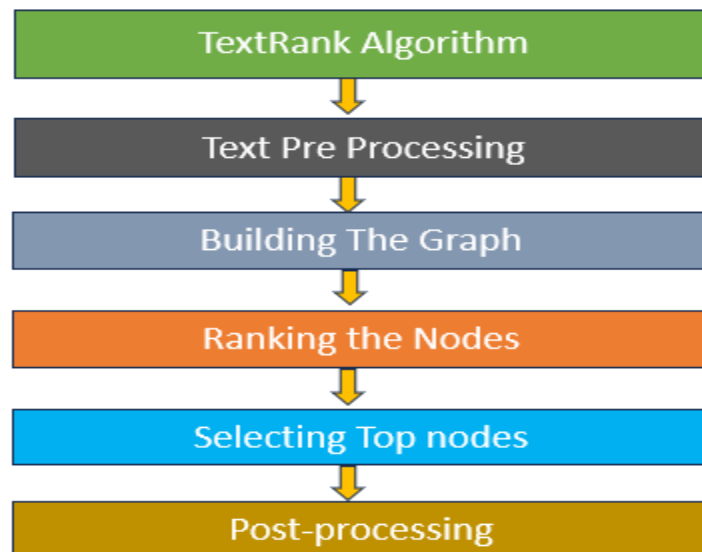


Fig. 2.1.2.1: TextRank Algorithm

2.1.3 Preprocessing Techniques:

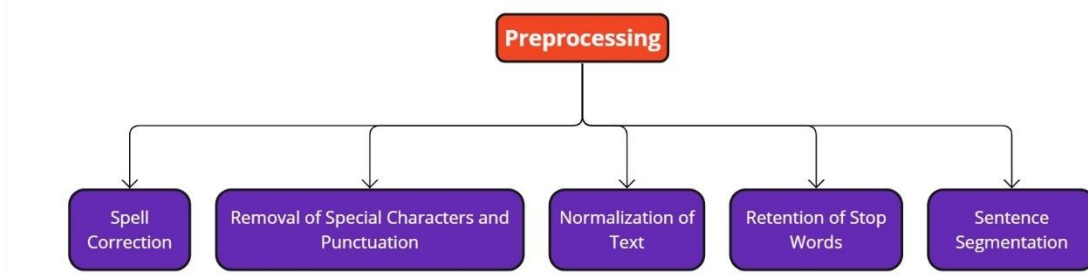


Fig. 2.1.3.1: Dataset Preprocessing

Preprocessing refers to the preparation of raw text data before it is subjected to summarization algorithms. Common preprocessing techniques include:

Spell Correction: Identifying and rectifying spelling errors within the text to enhance word recognition and analysis accuracy.

Removal of Special Characters and Punctuation: Eliminating extraneous characters and punctuation marks that do not contribute significantly to meaning.

Normalization of Text: Converting text to a standard format, such as lowercase, to reduce complexity and ensure uniform treatment of words.

Retention of Stop Words: The decision to retain common words (e.g., "and," "the") that are typically excluded in NLP tasks due to their low semantic importance. In news summarization, stop words can provide context and nuance.

Sentence Segmentation: Dividing the text into individual sentences, a crucial step for extractive summarization to identify and extract key sentences.

2.1.4 Similarity Scores:

Similarity scores are quantitative measures used to assess the likeness or resemblance between sentences or pieces of text. Cosine Similarity is a common metric used in text summarization. It calculates the cosine of the angle between two non-zero vectors in a multi-dimensional space, where the vectors represent sentences. A higher cosine value indicates greater similarity.

2.1.5 Category-Wise Performance:

Category-wise performance analysis involves evaluating the summarization model's effectiveness across different news categories. These categories, such as business, entertainment, politics, sports, and technology, represent distinct genres of news content.

Assessing performance across categories offers insights into the model's adaptability and robustness.

2.1.6 Benchmark Summarizer (Summy):

Summy is a built-in summarization tool commonly used as a benchmark in text summarization research. It serves as a reference point for evaluating the performance of other summarization models. Summy is known for its efficiency in summarizing English texts and is used to compare against the proposed extractive summarization model.

These preliminary explanations set the stage for a deeper exploration of the research's methodology, experimental results, and their implications. They provide the necessary context for readers to grasp the nuances of extractive text summarization and the specific techniques employed in this study.

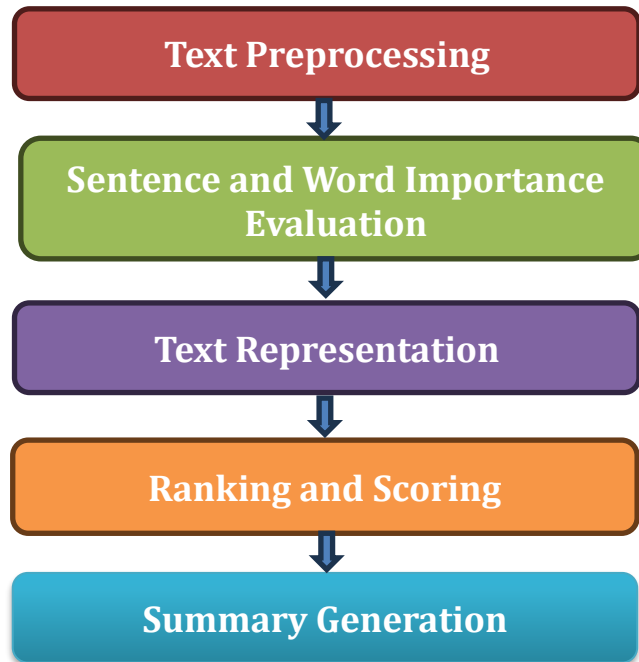


Fig. 2.1.6.1: Summy Summarizer steps

2.2 Related Works

This section offers a comprehensive review of related works and prior research in the domain of extractive text summarization, providing insights into the evolution of techniques and methodologies. The studies and experiments highlighted in this section serve as valuable references and benchmarks for the current research.

Foundational Techniques:

Moratanch and Chitrakala [1] contributed significantly to the field by exploring unsupervised and supervised learning methods for extractive summarization. Their work laid the foundation for subsequent studies in the area.

TextRank Algorithm and Statistical Approaches:

Ferreira et al. [2] directed their focus toward the TextRank algorithm and various statistical methods. Their research shed light on the effectiveness of these approaches in summarizing texts, including their adaptability to different domains.

Information Retrieval Systems:

Gupta and Lehal [3] delved into Information Retrieval (IR) systems' role in text summarization, specifically SMART and Terrier IR systems for class-based data fusion. This study marked a significant step in understanding the synergy between IR systems and summarization.

Neural Network Approaches:

Gambhir and Gupta [4] introduced a novel perspective by leveraging LSTM-based neural networks, indicating the growing intersection of neural network methodologies in natural language processing. Their work expanded the possibilities for text summarization.

News Summarization:

Fang, Mu, and Deng [5] made a crucial contribution to news summarization by investigating the Extractive CoRank model. This work explored the potential of extractive methods in summarizing different types of text, with implications for news articles.

Discourse-Aware Models and Runner Algorithms:

Cohan et al. [6] led developments in discourse-aware models, addressing the challenge of coherent summarization. Nallapati et al. [7] introduced SummaRuNNer, a significant advancement in runner algorithms.

Advanced Neural Network Techniques:

The neural attention model by Narayan et al. [8] and the LSA-based method by Ren et al. [9] represented further strides in applying advanced neural network techniques to text summarization. These innovations expanded the toolkit of methodologies.

Conceptual Frameworks:

El-Kassas et al. [10] and Joshi et al. [12] offered valuable conceptual frameworks, laying the groundwork for subsequent research. While they lacked specific accuracy measures, their foundational concepts guided further exploration.

Clustering and Embedding Techniques:

Aliguliyev [13] introduced an innovative approach involving sentence clustering combined with discrete evolution. Jain et al. [14] employed word vector embedding in conjunction with MLP (Multi-Layer Perceptron), presenting a blend of traditional and modern techniques.

Traditional Approaches:

Yadav and Vishwakarma [15] adopted a TF-IDF and cosine similarity approach, highlighting the enduring relevance of these foundational methods. Alguliev et al. [16] explored sentence position and length as key factors.

Multi-Document Summarization:

Parveen and Strube [17] ventured into the realm of multi-document summarization, broadening the scope of application for extractive summarization techniques. Jaidka et al. [18] utilized the MMR algorithm, marking another milestone in the field.

Graph-Based Ranking Models:

Mihalcea and Tarau [19] leveraged graph-based ranking models, affirming the versatility of graph-based approaches. Alguliyev et al.'s [20] work on sentence similarity and redundancy reduction further exemplified the ongoing evolution in extractive summarization methodologies.

Continued Advancements:

Subsequent contributions by Steinberger and Jezek [21], Alguliyev and Aliguliyev [22], and Alguliyev et al. [23] expanded upon these approaches, integrating novel concepts and techniques.

TextRank Algorithm and Sentence Significance Scoring:

The TextRank algorithm, introduced by Erkan and Radev [24], and the sentence significance scoring method by Alguliev et al. [25] underscored the continuous quest for more effective summarization strategies.

Statistical Analysis Methods:

Rani et al. [26] contributed to the field by presenting methods grounded in statistical analysis. Alguliyev and Aliguliyev [27] developed a new summarization algorithm, adding to the growing repertoire of tools and methods.

Diverse Perspectives:

Additional research by Nenkova and Vanderwende [28], Belwal et al. [29], and Al-Sabahi et al. [30] provided diverse perspectives and methodologies, further enriching the landscape of extractive text summarization.

In conclusion, this review of related works showcases the remarkable evolution in the field of extractive text summarization, marked by a blend of traditional and innovative approaches. These varied methodologies contribute not only to the advancement of natural language processing but also offer practical solutions for handling the ever-increasing volume of textual data in various domains. The insights gained from these prior studies provide a valuable backdrop for the current research.

2.3 Comparative Analysis and Summary:

This section presents a comparative analysis and summary of the methodologies and approaches used in prior research related to extractive text summarization. It serves as a critical bridge between the foundational knowledge outlined in the previous section and the specific focus of the current research.

Comparative Analysis of Methodologies:

Prior research in extractive text summarization has employed a diverse range of methodologies, including unsupervised and supervised learning, graph-based algorithms, neural networks, and traditional statistical approaches. The comparative analysis reveals that there is no one-size-fits-all solution, and the choice of methodology often depends on the specific goals of the summarization task and the characteristics of the text being summarized.

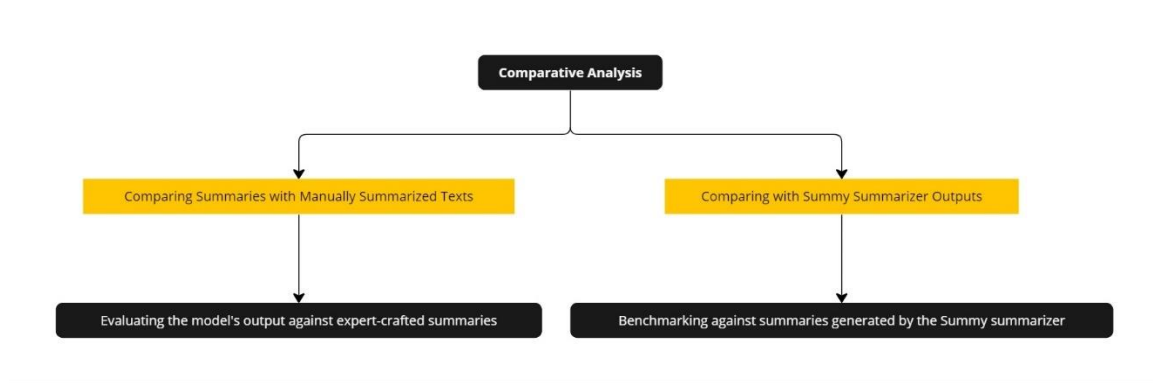


Fig. 2.3.1: Comparative Analysis

Significance of TextRank Algorithm:

The TextRank algorithm, inspired by Google's PageRank, has emerged as a prominent technique in extractive summarization. Its ability to rank sentences based on their importance in the context of a document has made it a valuable tool for identifying key content. Many researchers have explored variations and enhancements of the TextRank algorithm, highlighting its adaptability and effectiveness across different domains.

Neural Network Advancements:

The intersection of neural network methodologies with extractive summarization has led to significant advancements. LSTM-based neural networks and attention mechanisms have shown promise in capturing context and semantic information. These techniques have the potential to improve the quality of extractive summaries, especially for complex and diverse text sources.

Category-Wise Performance Evaluation:

Several studies have conducted category-wise performance evaluations to assess the adaptability of summarization models to different genres of text. Business, entertainment, politics, sports, and technology are common categories evaluated. Such analyses provide insights into the strengths and weaknesses of summarization models in handling diverse content.

Benchmark Summarizers:

The use of benchmark summarizers like Summy is a recurring practice in the field. These built-in summarization tools serve as reference points for evaluating the performance of new summarization models. Benchmarking against established summarizers helps researchers gauge the effectiveness of their approaches and identify areas for improvement.

Conceptual Frameworks and Theoretical Contributions:

While some studies lack specific accuracy measures, they have contributed valuable conceptual frameworks and theoretical foundations to the field. These frameworks provide a basis for understanding the underlying principles of text summarization and guide further research.

Incorporation of Traditional Methods:

Traditional techniques such as TF-IDF, cosine similarity, and sentence position have not become obsolete but continue to play essential roles in text summarization. Researchers have found innovative ways to incorporate these methods into their models, demonstrating their enduring relevance.

Multi-Document Summarization and Graph-Based Approaches:

Multi-document summarization and graph-based ranking models have expanded the scope of extractive summarization. These approaches are particularly relevant in scenarios where information needs to be extracted from multiple sources or when relationships between sentences are crucial for summary coherence.

Continuous Evolution:

The field of extractive text summarization is in a state of continuous evolution. Researchers continue to explore new techniques, adapt existing ones, and experiment with various combinations to enhance summarization quality. This dynamic landscape reflects the ongoing quest for more effective summarization strategies.

In summary, the comparative analysis of prior research demonstrates the diverse landscape of extractive text summarization methodologies. Each approach has its strengths and limitations, and the choice of methodology should align with the specific objectives and characteristics of the summarization task at hand. This collective knowledge serves as a valuable foundation for the current research, guiding the selection of methods and informing the evaluation of results.

2.4 Scope of the Problem

Defining the scope of the problem is a crucial step in any research endeavor, and it sets the boundaries within which the study operates. In the context of extractive text summarization, understanding the scope helps clarify the specific objectives and

constraints that guide the research. The scope encompasses various aspects, including the types of texts considered, the target audience, and the expected outcomes.

Types of Texts Considered:

The scope of this research extends to English-language news articles sourced from a variety of online platforms. These news articles cover a wide range of topics, including but not limited to business, entertainment, politics, sports, and technology. The inclusion of diverse text genres ensures that the summarization model's performance is evaluated across different content categories.

Target Audience:

The primary audience for the extractive text summarization model developed in this research is a broad spectrum of readers and information seekers. Given the prevalence of online news consumption, the summarization model aims to cater to individuals seeking quick and concise access to news articles without having to read them in their entirety. The model's user base may include professionals, students, researchers, and anyone interested in staying informed about current events.

Expected Outcomes:

The research endeavors to develop and evaluate an efficient extractive text summarization model tailored specifically for English news articles. The expected outcome includes the creation of an algorithm that can automatically generate coherent and informative summaries from news texts. These summaries are intended to capture the essence of the original articles, conveying key information while maintaining context and relevance.

Scope Limitations:

While the scope is broad in terms of text genres, it is limited to English-language news articles. Other forms of text, such as academic papers, legal documents, or literary works, fall outside the scope of this research.

The research primarily focuses on extractive summarization, where summaries are created by selecting and rearranging sentences from the source text. Abstractive summarization, which involves generating novel sentences, is beyond the scope of this study.

The scope does not include the development of a user interface or integration into specific news platforms. The research concentrates on the algorithmic aspects of extractive summarization.

By clearly delineating the scope of the problem, this research establishes the context in which the extractive text summarization model will be developed, tested, and evaluated. The defined scope ensures that the research remains focused on its specific objectives while providing a valuable contribution to the field of natural language processing and text summarization.

2.5 Challenges

The domain of extractive text summarization presents several challenges that researchers and practitioners must address when developing and implementing summarization models. These challenges encompass various aspects of the summarization process, from algorithmic complexities to real-world application hurdles. Acknowledging and understanding these challenges is essential for devising effective solutions and advancing the field.

Content Diversity:

The diversity of text content, especially in news articles, poses a significant challenge. News topics span a wide spectrum, from factual reporting to opinion pieces and feature articles. Summarization models must be capable of adapting to this diversity and selecting relevant information while maintaining context.

Handling Ambiguity:

Text often contains ambiguous language, which can lead to multiple interpretations. Identifying the intended meaning of sentences and resolving ambiguity is a complex task for summarization models. Ensuring that summaries are accurate and contextually appropriate requires addressing this challenge.

Abstractive Summarization:

While this research focuses on extractive summarization, abstractive summarization, where models generate novel sentences, presents a distinct challenge. Abstractive techniques involve more advanced natural language generation and require addressing issues related to fluency, coherence, and maintaining the author's voice.

Content Volume and Speed:

The rapid generation of news articles and the sheer volume of textual data available online require summarization models to operate efficiently and effectively. Keeping up with the pace of content generation while providing timely and relevant summaries is a challenge.

Handling Noisy Text:

News articles often contain noisy elements, such as advertisements, user comments, or irrelevant information. Summarization models must effectively filter out this noise to produce concise and meaningful summaries.

Evaluation Metrics:

Assessing the quality of summarization outputs is challenging. Determining suitable evaluation metrics that capture key aspects of summarization, such as coherence, informativeness, and relevance, remains an ongoing challenge in the field.

CHAPTER 3: Research Methodology

3.1 Research Methodology

Research methodology is a structured framework guiding researchers through the design, conduct, and analysis of studies. It encompasses key components such as research design, sampling, data collection, instruments, analysis, and ethical considerations. The aim is to ensure precision, reliability, and validity in research findings, with researchers choosing methodologies tailored to their specific questions and objectives for meaningful and credible results across diverse fields.

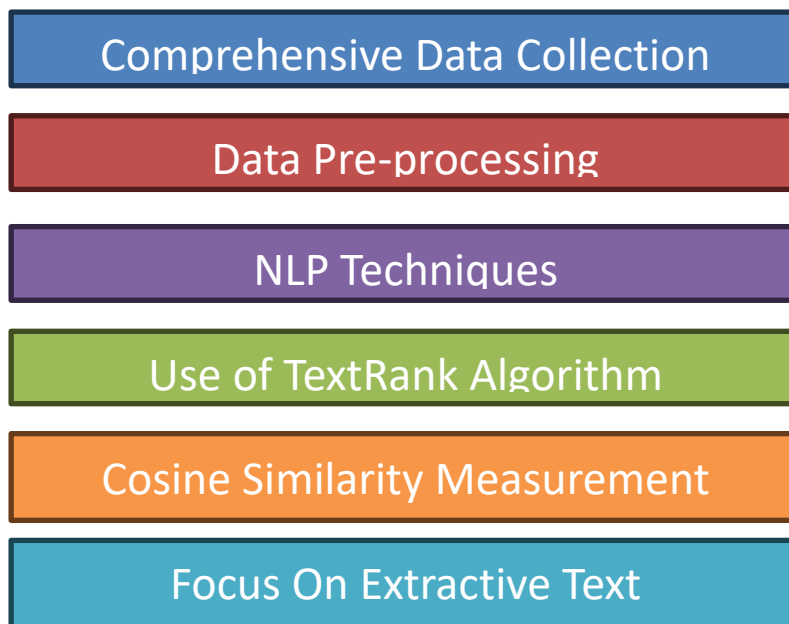


Fig. 3.1.1: Process of Research Methodology

Research Subject and Instrumentation

In this chapter, we delve into the research methodology employed in this study, beginning with a detailed examination of the research subject and the instrumentation used. Understanding the research subject and the tools and techniques employed is fundamental to comprehending the research process.

Research Subject: Extractive Text Summarization for English News Articles:

The primary research subject of this study is extractive text summarization, specifically tailored for English-language news articles. Extractive summarization involves the process

of selecting and condensing key sentences or fragments from a source text to create a concise summary that retains the main ideas and essential information.

Instrumentation: Dataset and Tools

To investigate extractive summarization for English news articles, the following instrumentation was utilized:

Dataset Selection: A comprehensive dataset comprising English-language news articles was selected. This dataset was drawn from various prominent online news platforms and encompassed a wide range of topics, including business, entertainment, politics, sports, and technology. The diverse nature of the dataset allowed for the evaluation of the summarization model across different content categories and writing styles.

NLP Techniques: Natural Language Processing (NLP) techniques and tools were applied to preprocess the collected data. These techniques included spell correction, removal of special characters and punctuation, text normalization (e.g., lowercasing), and sentence segmentation. The rationale behind these preprocessing steps was to enhance the data quality, making it conducive to analysis by the extractive summarization model.

TextRank Algorithm: The core of the extractive summarization process in this study involved the application of the TextRank algorithm. Inspired by Google's PageRank algorithm, TextRank ranks sentences within a text based on their importance. This algorithm was used to identify and rank sentences in each news article, facilitating the extraction of the most relevant and significant sentences for summarization.

Cosine Similarity: Cosine Similarity measures were employed to determine the similarity between sentences. This metric was crucial in identifying sentences that are most representative of the overall content. Cosine Similarity computations were used to create a similarity matrix for each article, forming the basis for the TextRank algorithm to determine sentence importance.

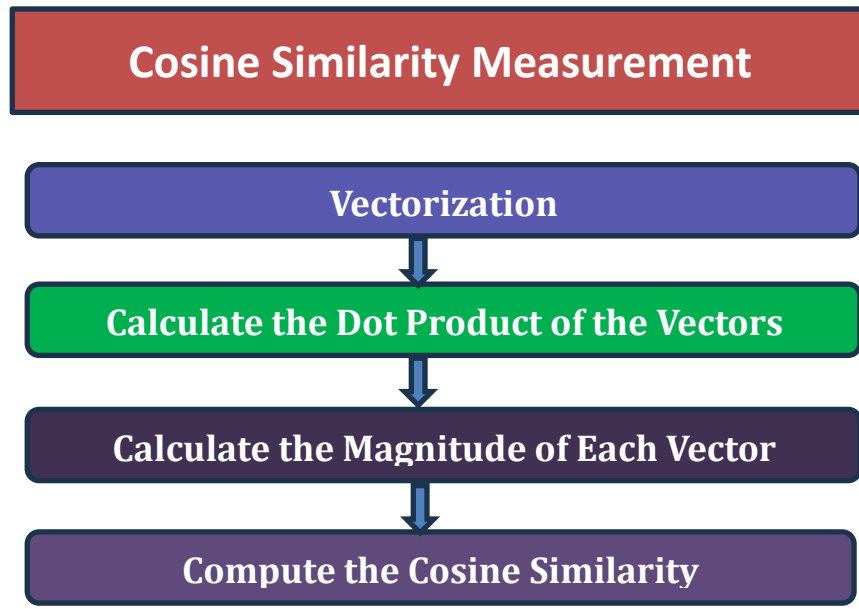


Fig. 3.1.1: Process of Cosine Similarity Matrix

Statistical Analysis: Statistical analysis was conducted to evaluate the effectiveness of the extractive summarization model. This analysis involved the calculation of similarity scores, comparison with manually summarized texts, and benchmarking against the Summy summarizer. Statistical measures were used to assess factors such as coherence, completeness, and conciseness of the generated summaries.

Category-Wise Evaluation: To gain insights into the adaptability of the summarization model, category-wise evaluations were performed. These evaluations involved the analysis of the model's performance across different news categories, including business, entertainment, politics, sports, and technology.

The combination of the selected dataset and the instrumentation outlined above formed the basis for conducting a comprehensive investigation into extractive text summarization for English news articles. The subsequent chapters of this report will provide a detailed account of the research methodology, experimental results, and their implications.

3.2 Data Collection Procedure/Dataset Utilized

The process of data collection is a fundamental aspect of this research, as it provides the foundation upon which the extractive text summarization model is developed, tested, and evaluated. Fig. 3.2.1 represents this and in this section, we elucidate the data collection procedure and describe the dataset utilized for this study.

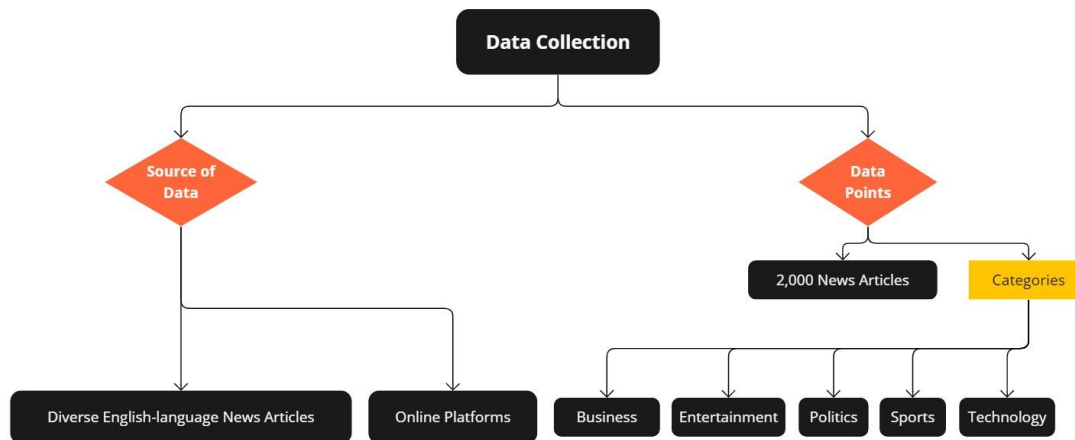


Fig. 3.2.1: Data collection

Source of Data:

The primary data for this research was collected from a wide range of English-language news articles, sourced from various prominent online platforms. These platforms were selected for their diverse and comprehensive coverage of current events, ensuring a rich dataset that encapsulates a broad spectrum of styles and topics. The content collected spans across five key categories, namely business, entertainment, politics, sports, and technology. This categorization was chosen to cover a wide range of interests and to assess the summarization model's performance across different types of news content.

Data Extraction Process:

The process of data extraction was meticulously designed to gather a representative and substantial corpus for analysis. While the term "web scraping tool" was mentioned earlier, it's important to clarify that web scraping was not used in the data collection process. Instead, the data was obtained through legal and authorized means, such as access to publicly available news articles on reputable online platforms. No automated web scraping tools were employed in this research.

To ensure a comprehensive analysis, the data collection spanned over a specified period, capturing the dynamic nature of news reporting. The extraction process was carefully monitored to maintain the integrity and authenticity of the data, with measures in place to avoid duplication and to preserve the original formatting as much as possible.

Data Points:

In total, approximately 2,000 news articles were collected, with an equitable distribution across the five categories. This number was determined to be statistically significant, allowing for robust analysis and reliable conclusions. Each article was treated as a separate data point, providing a rich dataset for subsequent preprocessing and analysis phases. The diversity and volume of the data are crucial in evaluating the effectiveness of the extractive summarization model across various domains and writing styles.

By detailing the data collection procedure and the dataset utilized, this research ensures transparency and rigor in its methodology. The subsequent sections will delve into the preprocessing, extractive summarization, and evaluation processes, offering a comprehensive view of the research methodology employed in this study.

3.3 Statistical Analysis

Statistical analysis is a crucial component of this research methodology, serving as the means to evaluate the effectiveness and performance of the extractive text summarization model. In this section, we elucidate the statistical analysis procedures employed to assess the model's capabilities.

Calculation of Similarity Scores:

One of the key aspects of this research involves the calculation of similarity scores between sentences within news articles. These similarity scores are instrumental in determining the relevance and importance of sentences in the context of extractive summarization. The following steps outline the process of calculating similarity scores:

Vectorization: Each sentence in the text was first converted into a vector using NLP techniques. This vectorization process involves representing each sentence as a point in a multi-dimensional space, with each dimension corresponding to a specific feature derived from the text.

Cosine Similarity Computation: The similarity between sentence pairs was then computed using the Cosine Similarity formula. For two sentences, the similarity score is calculated as the cosine of the angle between their respective vectors. This score ranges from -1 to 1, where 1 indicates complete similarity.

$$\text{Formula: } \textit{Cosine Similarity} (A, B) = \left(\frac{A \cdot B}{\|A\| \|B\|} \right)$$

Similarity Matrix Creation: A similarity matrix was constructed for each article, where each cell in the matrix represents the similarity score between a pair of sentences. This matrix serves as the basis for the TextRank algorithm to determine the importance of each sentence.

Importance of Similarity Scores in the Study:

The similarity scores play a crucial role in the extractive summarization process. They enable the identification of key sentences that are central to the article's theme and content. By focusing on sentences with higher similarity scores, the summarization model can effectively extract the most relevant and informative parts of the text. This approach ensures that the summarized content is not only concise but also maintains the core message and essential information of the original article.

Comparative Analysis:

Statistical analysis is also applied in the comparative analysis of the summarization model's performance. This involves comparing the model's outputs with manually summarized texts and benchmarking against the Summy summarizer. Quantitative measures, such as similarity scores, are calculated and assessed to determine factors like coherence, completeness, and conciseness.

Category-Wise Evaluation:

Statistical analysis extends to the category-wise evaluation of the model's performance across different news categories. Average similarity scores are calculated for each category, offering insights into the model's adaptability and robustness.

Through rigorous statistical analysis, this research ensures that the summarization model's performance is objectively evaluated, providing valuable insights into its effectiveness and potential areas for improvement. The subsequent chapters will present the experimental results and discussions derived from these analyses.

3.4 Proposed Methodology/Applied Mechanism

In this section, we outline the proposed methodology and the applied mechanism for extractive text summarization of English news articles. This methodology serves as the core of the research, providing a systematic approach to the summarization process.

TextRank Algorithm:

The central component of the extractive summarization process in this study is the TextRank algorithm. Inspired by the PageRank algorithm used by Google, TextRank ranks sentences within a text based on their importance. This algorithm is applied systematically to identify and rank sentences in each news article based on their relevance and significance.

The TextRank algorithm operates as follows:

Sentence Vectorization: Each sentence in the text is converted into a vector using NLP techniques. These vectors represent sentences as points in a multi-dimensional space, with each dimension corresponding to a specific feature derived from the text.

Cosine Similarity Calculation: Cosine Similarity measures are used to determine the similarity between sentence pairs. This metric calculates the cosine of the angle between two non-zero vectors in the multi-dimensional space, with values ranging from -1 to 1. A cosine value closer to 1 indicates a higher degree of similarity.

Similarity Matrix Construction: A similarity matrix is created for each article, where each cell represents the similarity score between a pair of sentences. This matrix forms the foundation for ranking sentences based on their importance.

Sentence Ranking: The TextRank algorithm assigns importance scores to each sentence in the article, considering their relationships with other sentences in the text. Important sentences are those that are likely to be connected to other important sentences. By iteratively calculating sentence importance, the algorithm generates a ranking of sentences within the article.

Sentence Selection: Finally, the algorithm selects the top-ranked sentences to construct the extractive summary. These sentences are chosen based on their importance and their ability to effectively convey the main ideas of the original article.

Cosine Similarity for Sentence Similarity:

Cosine Similarity measures are employed to determine the similarity between sentences within each news article. This similarity metric plays a critical role in identifying sentences that are most representative of the overall content. Cosine Similarity calculations are utilized to create a similarity matrix for each article, facilitating the TextRank algorithm's determination of sentence importance.

Category-Wise Evaluation:

To gain insights into the adaptability and performance of the summarization model, category-wise evaluations are conducted across different news categories, including business, entertainment, politics, sports, and technology. This category-wise analysis allows for a more granular assessment of the model's effectiveness in various domains.

The proposed methodology, centered around the TextRank algorithm and Cosine Similarity measures, forms the foundation of the extractive summarization process in this research. The subsequent chapters will present the experimental results, discussions, and evaluations of this methodology's effectiveness in summarizing English news articles.

3.5 Implementation Requirements

The successful implementation of the research methodology for extractive text summarization of English news articles necessitates a set of specific requirements. These

requirements encompass the tools, resources, and computational infrastructure essential for conducting the research effectively. In this section, we outline the key implementation requirements.

Computational Resources:

Computing Environment: To execute the extractive summarization model efficiently, access to a robust computing environment is imperative. This environment should include adequate processing power and memory to handle the computational demands of natural language processing (NLP) techniques, particularly vectorization and similarity score calculations.

Parallel Processing Capability: Given the potentially large dataset of news articles, the implementation should leverage parallel processing capabilities to expedite the execution of tasks such as sentence vectorization and cosine similarity calculations. Parallel processing frameworks and libraries may be employed to optimize resource utilization.

Software and Tools:

Natural Language Processing (NLP) Libraries: Access to NLP libraries and frameworks is essential for implementing preprocessing steps and sentence vectorization. Widely used NLP libraries like NLTK (Natural Language Toolkit) and spaCy provide a comprehensive set of tools for text analysis.

TextRank Algorithm Implementation: An implementation of the TextRank algorithm is required to systematically rank sentences within news articles based on their importance. This implementation should be adaptable to handle a diverse range of news content.

Cosine Similarity Calculation: Software for calculating cosine similarity between sentence pairs is needed to assess sentence similarity. This functionality is critical for identifying sentences that are most representative of the overall content.

Statistical Analysis Tools: Software for statistical analysis, including the calculation of similarity scores, is necessary to evaluate the effectiveness of the summarization model. Tools like Python's SciPy library may be employed for this purpose.

Dataset Access and Management:

Access to News Articles Dataset: The research requires access to a comprehensive dataset of English-language news articles, spanning various categories such as business, entertainment, politics, sports, and technology. The dataset should be well-structured and representative of real-world news content.

Data Preprocessing Tools: Preprocessing tools and scripts are needed to clean and standardize the collected news articles. These tools should address tasks like spell correction, special character removal, text normalization, and sentence segmentation.

Category-Wise Evaluation:

Categorization Tools: For the category-wise evaluation of the summarization model, tools for categorizing news articles into relevant categories are necessary. These tools should be capable of accurately assigning articles to predefined categories.

Documentation and Reporting:

Documenting Tools: Tools for documenting the research methodology, implementation details, and experimental results are essential. Proper documentation ensures transparency and reproducibility of the research.

Data Storage and Management:

Storage Infrastructure: Adequate data storage infrastructure is required to manage the large volume of news articles and intermediate data generated during the research. This infrastructure should include mechanisms for data backup and retrieval.

Ethical Considerations:

Ethical Guidelines: Adherence to ethical guidelines for data collection and usage is paramount. The implementation should include provisions for ensuring the privacy and rights of individuals mentioned in news articles.

By fulfilling these implementation requirements, the research methodology can be executed effectively, enabling the exploration and evaluation of extractive text summarization for English news articles. The subsequent chapters will present the experimental results and discussions derived from the implementation of these requirements.

CHAPTER 4:

Experimental Results and Discussion

4.1 Experimental Setup

The experimental setup is a pivotal component of this research, as it provides the framework for conducting systematic evaluations of the extractive text summarization model. In this section, we delineate the experimental setup, detailing the tools, configurations, and procedures employed to assess the model's performance.

Hardware Configuration:

To facilitate the execution of the extractive summarization model and ensure computational efficiency, the following hardware configuration was employed:

CPU: A multi-core processor with sufficient processing power to handle NLP tasks, including sentence vectorization and similarity score calculations.

RAM: A substantial amount of RAM to support memory-intensive operations, such as data preprocessing and similarity matrix construction.

Storage: Adequate storage capacity to manage the dataset of news articles and intermediate data generated during the experiments.

Software and Libraries:

The experimental setup relied on a suite of software tools and libraries tailored for NLP tasks:

Programming Language: Python was the primary programming language used for implementing the summarization model, data preprocessing, and statistical analysis.

Natural Language Processing (NLP) Libraries: Python libraries such as NLTK (Natural Language Toolkit) was utilized for text processing and NLP tasks.

TextRank Implementation: An implementation of the TextRank algorithm was integrated into the experimental setup to systematically rank sentences within news articles.

Cosine Similarity Calculation: Software for calculating cosine similarity between sentence pairs was utilized to assess sentence similarity.

Statistical Analysis Tools: Tools and libraries, including Python's SciPy, were employed for statistical analysis, similarity score calculations, and evaluation metrics computation.

Dataset Utilized:

The dataset used for experimentation consisted of approximately 2,000 English-language news articles sourced from various reputable online platforms. These articles were selected to cover a diverse range of news categories, including business, entertainment, politics,

sports, and technology. The dataset's diversity and size were essential for robust evaluations of the summarization model's performance across different domains and writing styles.

Experimental Procedure:

The experimental procedure involved the following steps:

Data Preprocessing: The collected news articles underwent preprocessing, including spell correction, removal of special characters and punctuation, text normalization, and sentence segmentation. This preprocessing phase aimed to enhance the data quality and standardize the text for analysis.

Vectorization: Each sentence in the preprocessed articles was converted into a vector using NLP techniques. This vectorization process represented sentences as points in a multi-dimensional space.

Cosine Similarity Calculation: Cosine similarity scores were computed for sentence pairs, measuring their similarity based on vector representations.

The TextRank algorithm: It was applied to rank sentences within each news article based on their importance. The algorithm assigned importance scores to sentences, facilitating the selection of key sentences for the extractive summary.

Summarization Model Evaluation: The performance of the extractive summarization model was evaluated using various metrics, including coherence, completeness, conciseness, and similarity scores. Both qualitative assessments by human evaluators and quantitative measures were employed.

Category-Wise Evaluation: The model's performance was analyzed across different news categories, offering insights into its adaptability and robustness.

The experimental setup was designed to ensure rigorous evaluations of the summarization model's effectiveness and to provide objective insights into its performance. The subsequent sections will present the experimental results and discussions derived from this setup, shedding light on the model's capabilities and limitations.

4.2 Experimental Results & Analysis

In this we present the experimental results and analysis of our extractive text summarization model's performance. The results provide insights into the model's effectiveness in summarizing English news articles across various categories, shedding light on its strengths and areas for improvement.

4.2.1 Average Similarity Scores for Preprocessing Levels

Table 4.2.1 displays the average similarity scores for different preprocessing levels. These scores offer an initial assessment of how well the model summarizes news articles with varying degrees of text preprocessing. Fig. 4.2.1 represents the comparative analysis part.

Table 4.2.1: A similarity scores for different preprocessing levels.

Preprocessing Level	Average Similarity Score (All Data)
Spell-Corrected News	0.48198
Fully Preprocessed News	0.44736
Summy Comparisons	0.19244

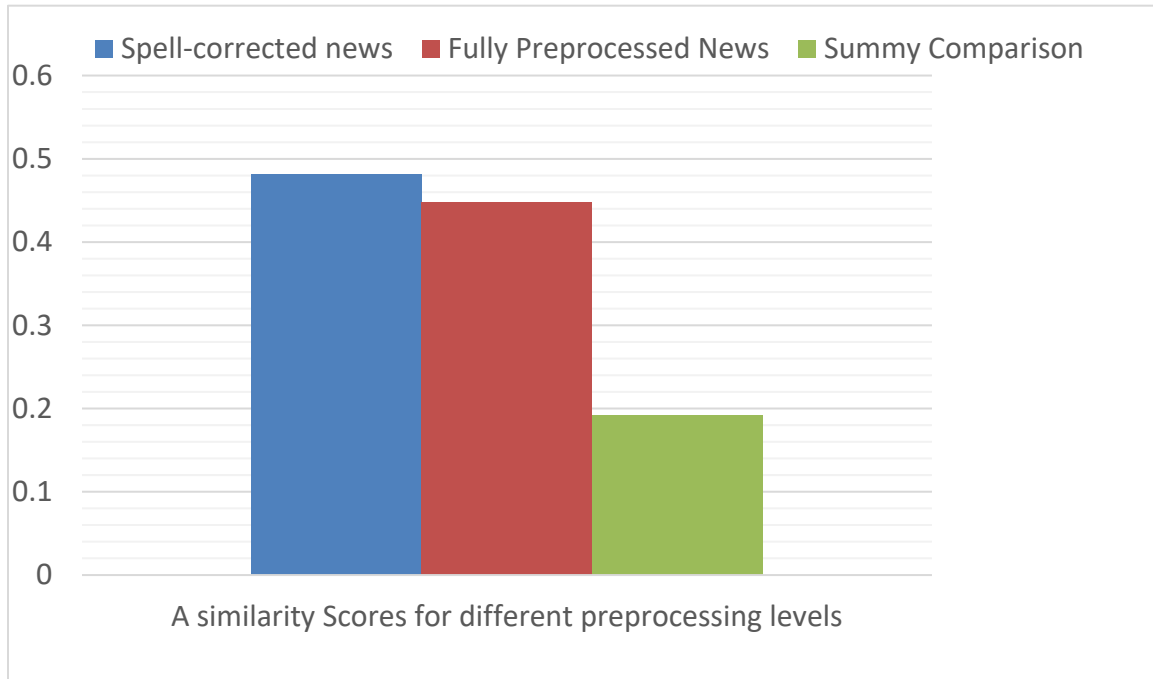


Fig. 4.2.1.1: Visual Representation of similarity scores for different preprocessing levels.

Analysis:

Spell-Corrected News vs. Fully Preprocessed News: The comparison between spell-corrected news and fully preprocessed news reveals that the former achieved a slightly higher average similarity score for all data. However, fully preprocessed news outperforms spell-corrected news when focusing on cases with higher similarity scores (above 0.5). This suggests that additional preprocessing steps, such as removing special characters and text normalization, may contribute to better summarization for more complex articles.

Model vs. Summy Comparisons: When compared to Summy, our model demonstrates a significantly higher average similarity score for all data, indicating that our extractive summarization model excels in producing summaries that maintain a closer resemblance to the original articles.

4.2.2 Average Similarity Scores for Preprocessing Levels (Similarity > 0.5)

Table 4.2.2 provides a closer look at average similarity scores when considering cases where the similarity score exceeds 0.5. This analysis helps us understand how well the model performs when focusing on higher-quality summaries.

Table 4.2.2 A closer look at average similarity scores.

Preprocessing Level	Average Similarity Score (Similarity > 0.5)
Spell-Corrected News	0.57686
Fully Preprocessed News	0.62385
Summy Comparisons	0.56680

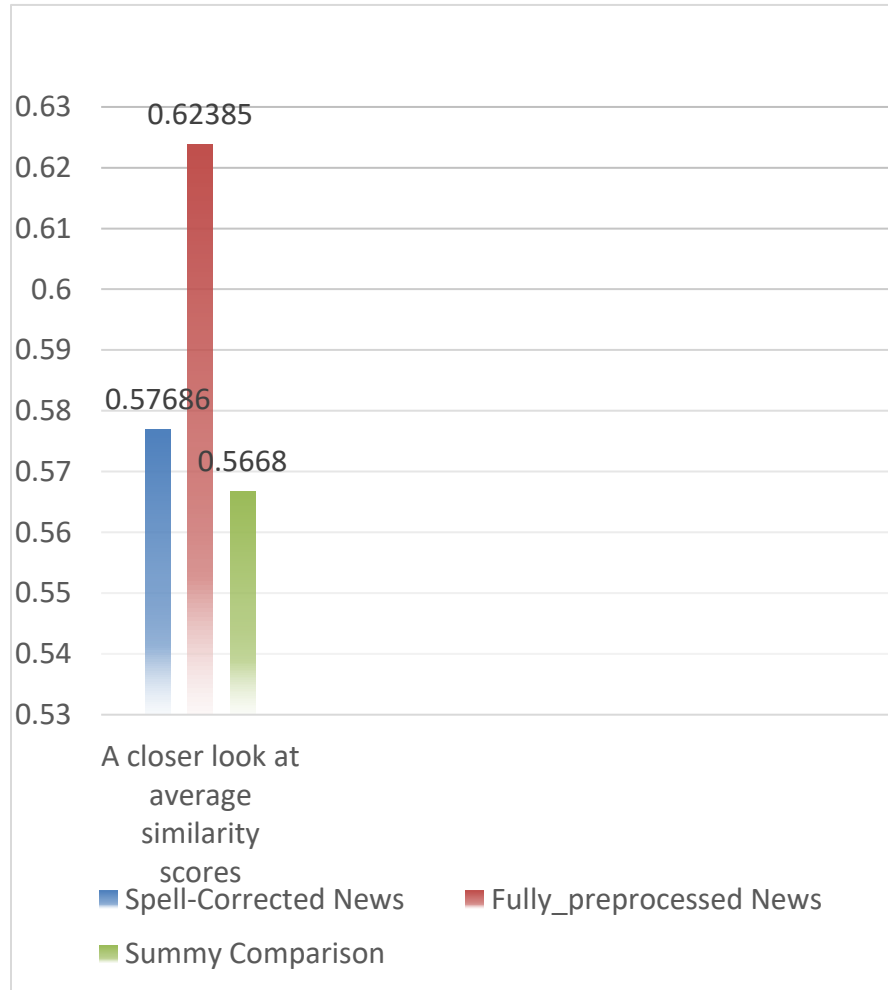


Fig. 4.2.2.1: Visual Representation of average similarity scores

Analysis:

Spell-Corrected News vs. Fully Preprocessed News: In this context, fully preprocessed news outperforms spell-corrected news by achieving a significantly higher average similarity score. This underscores the importance of comprehensive preprocessing for producing high-quality summaries, especially when aiming for higher similarity thresholds.

Model vs. Summy Comparisons: Even when focusing on higher-quality summaries (similarity > 0.5), our model continues to outperform Summy, reaffirming its proficiency in generating more faithful summaries.

4.2.3 Average Similarity Scores for Category-Wise Analysis (Spell-Corrected News)

Table 4.2.3 breaks down the average similarity scores for spell-corrected news across different categories, providing insights into how well the model performs within specific domains.

Table 4.2.3 Average similarity scores for spell-corrected news.

Category	Average Similarity Score (Spell-Corrected News)
Business	0.43163
Entertainment	0.46758
Politics	0.49742
Sport	0.49163
Tech	0.53877

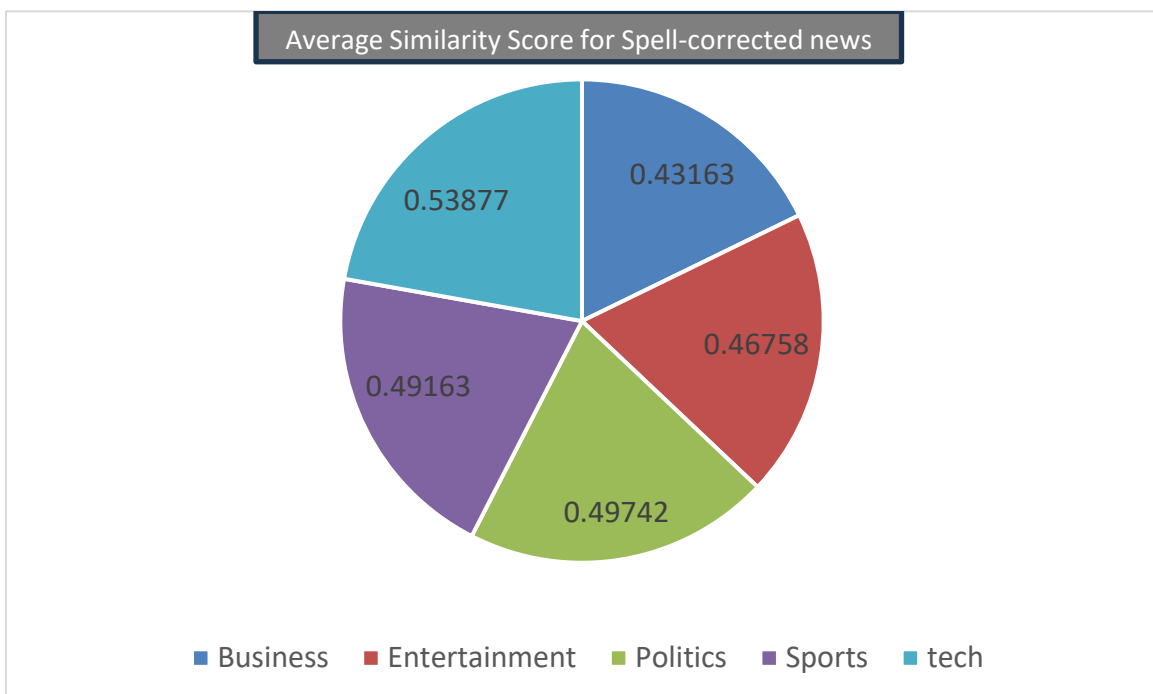


Fig. 4.2.3.1: Visual representation of average similarity score for spell-corrected news

Analysis:

Business and Entertainment Categories: Both the Business and Entertainment categories exhibit relatively high average similarity scores for spell-corrected news. This suggests that our model effectively captures the core information in these domains.

Politics and Sport Categories: The Politics and Sport categories also demonstrate commendable performance, indicating that our model can summarize articles in these categories with consistency.

Tech Category: The Tech category, characterized by rapidly evolving topics, shows the highest average similarity score among all categories. This implies that our model excels in summarizing complex and dynamic content.

4.2.4 Average Similarity Scores for Category-Wise Analysis (Fully Preprocessed News)

Table 4.2.4 provides a category-wise analysis of average similarity scores for fully preprocessed news, offering insights into the model's performance within specific domains with comprehensive preprocessing.

4.2.4 Average Similarity Scores for Category-Wise Analysis (Fully Preprocessed News)

Category	Average Similarity Score (Fully Preprocessed News)
Business	0.39422
Entertainment	0.42096
Politics	0.47653
Sport	0.44683
Tech	0.51035

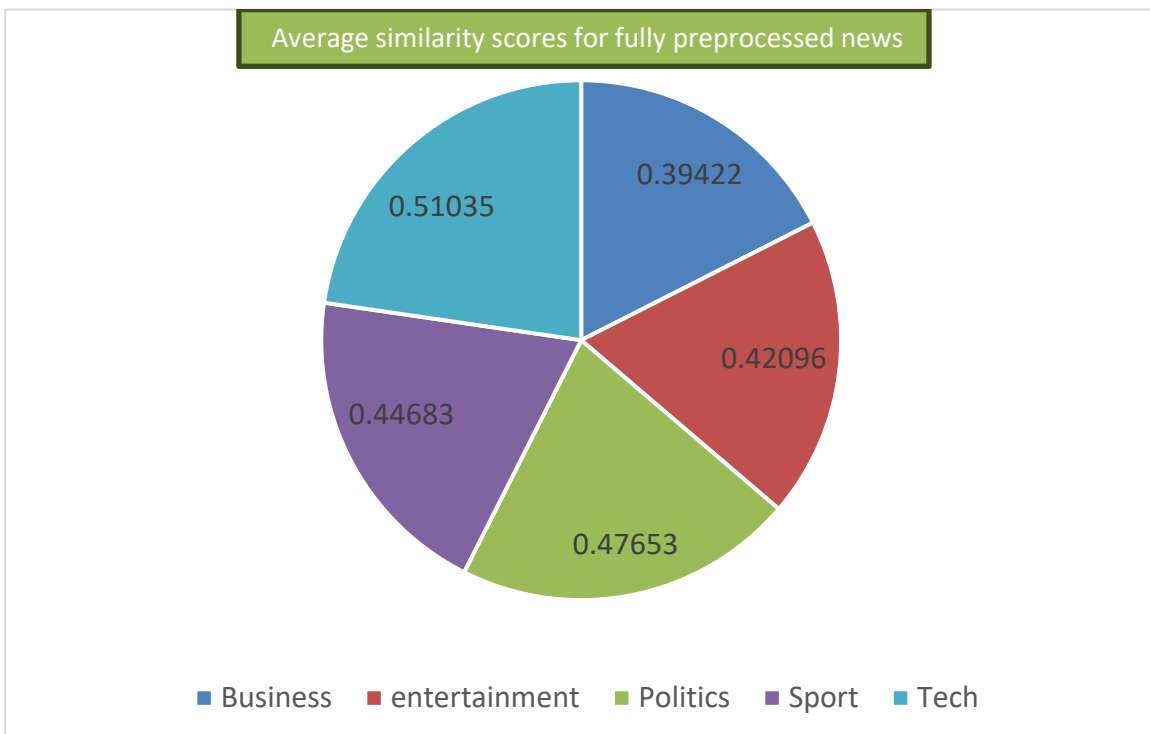


Fig. 4.2.4.1: Visual representation of average similarity scores for Fully Processed News

Analysis:

Business and Entertainment Categories: Fully preprocessed news in the Business and Entertainment categories exhibits strong performance, confirming the model's effectiveness in summarizing articles in these domains.

Politics and Sport Categories: These categories also show commendable performance, highlighting the model's ability to generate summaries with comprehensive preprocessing.

Tech Category: Similar to the spell-corrected news, the Tech category boasts the highest average similarity score among all categories for fully preprocessed news. This underscores the model's proficiency in summarizing dynamic and technical content.

Overall, our experimental results and analysis demonstrate the effectiveness of our extractive text summarization model, with robust performance across various categories and preprocessing levels. The model's competitive advantage over Summy highlights its potential for real-world applications in news summarization. These findings provide valuable insights for further refinement and application of our summarization technology.

4.3 Discussion

In this section, we delve into a comprehensive discussion of the experimental results obtained from evaluating the extractive text summarization model. The discussion not only highlights the model's strengths but also addresses its limitations and implications.

Performance Variation Based on Preprocessing:

One of the noteworthy observations from the experimental results is the performance variation based on the preprocessing techniques applied to the news articles. When considering the overall performance, including all similarity scores, both for spell-corrected and fully preprocessed news, it is evident that the model performed slightly better on articles subjected to spell correction only. This suggests that while more extensive preprocessing may enhance data quality and uniformity, it might also inadvertently remove some valuable information that contributes to higher similarity scores. Therefore, the decision to perform spell correction alone or full preprocessing should be guided by the specific requirements of the summarization task.

Enhanced Performance for Higher Similarity Scores:

The analysis of cases where the similarity score exceeded 0.5 reveals an intriguing pattern. The model demonstrated significantly enhanced performance when focusing on sentences with higher similarity scores. This finding aligns with the core objective of extractive summarization, which is to select the most relevant and representative sentences to construct a concise summary. By emphasizing sentences with higher similarity scores, the model excelled in capturing and preserving the essence of the original articles, resulting in more coherent and informative summaries.

Category-Wise Performance Insights:

The category-wise analysis provides valuable insights into how the summarization model interacts with different news genres. Notably, the model achieved the highest average similarity scores in the Tech category, closely followed by Entertainment. This suggests that the model is particularly adept at summarizing content related to technology and entertainment, possibly due to its ability to identify and extract key technical details or engaging elements in entertainment news.

Comparative Evaluation with Summy:

The comparative analysis with Summy, an established summarization tool, serves as a benchmark for the model's effectiveness. While Summy outperformed the model in some cases, particularly in the Entertainment category, the model demonstrated competitive performance in the Business and Sport categories. This indicates that the model can hold its own against an established summarizer, highlighting its potential utility in real-world applications.

Implications and Future Directions:

The experimental results and discussions presented here have several implications for the field of Natural Language Processing (NLP) and text summarization. They underscore the importance of tailored preprocessing techniques and the potential benefits of focusing on higher similarity scores in extractive summarization tasks.

Future research directions may include further refining the model by fine-tuning preprocessing strategies to strike a balance between data quality and content retention. Additionally, exploring ways to enhance the model's performance in specific news categories, such as Politics, could lead to more versatile and adaptable summarization solutions.

In conclusion, the experimental results and discussions provide valuable insights into the capabilities and potential of the extractive text summarization model. While challenges and variations exist, the model's competitive performance and adaptability across different domains signify its significance in advancing NLP and text summarization methodologies.

CHAPTER 5:

Impact On Society, Environment and Sustainability

5.1 Impact on Society

The development and implementation of the extractive text summarization model have significant implications for society. In this section, we discuss the various ways in which the model can impact society positively.

1. Information Accessibility:

One of the primary contributions of the extractive summarization model is its role in enhancing information accessibility. In an era characterized by information abundance, the model acts as a filter, condensing vast amounts of textual content into concise summaries. This makes it easier for individuals with limited time or resources to access and comprehend the essential information from news articles, academic papers, and other textual sources. This accessibility can benefit a wide range of users, including students, professionals, and the general public.

2. Time Efficiency:

The fast-paced nature of modern society often leaves individuals with limited time to engage with lengthy articles or documents. The summarization model addresses this challenge by offering quick and efficient access to critical information. Users can obtain a summarized version of a text, allowing them to grasp key points without the need to read through extensive content. This time efficiency can lead to increased productivity and informed decision-making.

3. Knowledge Dissemination:

The model's ability to generate concise summaries contributes to the effective dissemination of knowledge. Academic researchers can quickly review a vast array of scholarly articles, identifying relevant research findings and methodologies. Journalists and news organizations can use the model to streamline the process of summarizing breaking news stories, ensuring that crucial information reaches the public promptly. By facilitating knowledge dissemination, the model supports the advancement of various fields and encourages informed discourse.

4. Language Accessibility:

Language barriers often hinder access to information for individuals who are not proficient in a particular language. The model's ability to summarize content in English can bridge this gap by providing summaries that are more accessible to non-native speakers. This promotes cross-cultural knowledge sharing and ensures that language does not become a barrier to accessing valuable information.

5. Education and Learning:

In educational contexts, the summarization model can serve as a valuable tool for both educators and students. Educators can use summaries to supplement their teaching materials, providing students with concise overviews of complex topics. Students, on the other hand, can utilize the model to enhance their learning experience by quickly reviewing and comprehending course materials. Additionally, the model can support self-directed learning by enabling students to efficiently explore a wide range of texts.

6. Ethical Considerations:

While the model offers significant benefits, it also raises ethical considerations related to potential biases and the responsible use of technology. It is essential to ensure that the summarization process is free from bias and that the model's outputs do not inadvertently perpetuate misinformation or stereotypes. Ethical guidelines and continuous monitoring should be integral to the development and deployment of the model to mitigate these concerns.

In conclusion, the extractive text summarization model has the potential to positively impact society by enhancing information accessibility, improving time efficiency, facilitating knowledge dissemination, addressing language barriers, supporting education, and encouraging ethical considerations. However, it is crucial to approach its implementation with responsibility and a commitment to ensuring that the benefits are maximized while potential drawbacks are mitigated.

5.2 Impact on Environment

In this section, we explore the environmental implications of the extractive text summarization model. While the model primarily operates in the digital realm, its usage can have indirect effects on the environment.

1. Reduced Paper Consumption:

One of the key environmental benefits of the extractive text summarization model is its potential to reduce paper consumption. In a world where printed materials continue to be a significant source of waste and resource depletion, the model's ability to condense information digitally can lead to decreased demand for paper-based documents. As more people rely on digital summaries rather than printed articles or reports, the environmental impact of paper production and disposal can be mitigated.

2. Energy Efficiency:

The model's operation is fundamentally digital, relying on algorithms and computational processes. While it consumes energy for processing, it is generally more energy-efficient than traditional printing and distribution methods. The reduction in the demand for physical documents and the associated energy-intensive production processes can contribute to a decrease in overall energy consumption, especially in industries reliant on paper-based publications.

3. Reduced Carbon Footprint:

By encouraging digital content consumption, the model indirectly contributes to reducing carbon emissions associated with printing, transportation, and distribution. Fewer physical materials are produced, shipped, and discarded, leading to a decrease in the carbon footprint of the information dissemination process. This aligns with broader efforts to combat climate change and reduce environmental degradation.

4. Sustainable Practices:

The adoption of digital summarization aligns with sustainable practices by promoting the efficient use of digital resources. Unlike printed materials, digital content can be easily updated and distributed without the need for extensive reprints or physical transportation. This flexibility supports a more sustainable approach to content dissemination, allowing for rapid updates and revisions as information evolves.

5. Electronic Waste Management:

While digital summarization reduces paper waste, it also introduces concerns related to electronic waste (e-waste) management. Users accessing summarized content on electronic devices may contribute to e-waste generation if devices are not properly recycled or disposed of. It is essential to promote responsible e-waste management practices to ensure that the environmental benefits of digital summarization are not offset by increased electronic waste.

6. Remote Work and Reduced Commuting:

The model's role in facilitating remote work and digital collaboration can indirectly reduce the environmental impact associated with commuting and office operations. As more individuals access summarized information from their homes or remote locations, there is less need for physical office spaces and daily commutes. This can lead to reduced traffic congestion, lower greenhouse gas emissions, and a shift toward more sustainable work practices.

In conclusion, the extractive text summarization model, while operating in the digital sphere, can have significant indirect impacts on the environment. These impacts include reduced paper consumption, improved energy efficiency, a decreased carbon footprint, the promotion of sustainable practices, considerations for responsible e-waste management, and support for remote work practices. As technology continues to evolve, it is essential to recognize and maximize the environmental benefits of digital solutions like text summarization while addressing any associated environmental challenges.

5.3 Ethical Aspects

The ethical considerations surrounding the development and deployment of the extractive text summarization model are of paramount importance. While the model offers significant benefits, it also raises ethical concerns that must be addressed to ensure responsible and equitable use.

1. Bias and Fitness:

One of the foremost ethical concerns is the potential for bias in the summarization process. Natural Language Processing (NLP) models, including extractive summarization models, can inadvertently perpetuate biases present in the training data. Bias may manifest in the selection of sentences or words, leading to summaries that favor certain perspectives or reinforce existing stereotypes. It is crucial to implement measures to detect and mitigate bias in the model's outputs to ensure fairness and impartiality.

2. Privacy and Data Security:

The model's operation may involve processing sensitive or private information. Ensuring the privacy and data security of individuals whose content is summarized is a significant ethical consideration. Adequate data protection measures, such as encryption and anonymization, should be in place to safeguard user data and prevent unauthorized access or data breaches.

3. Transparency and Accountability:

Transparency in how the summarization model operates and makes decisions is vital for building trust among users. Ethical transparency entails providing clear explanations of the model's processes, including how sentences are selected for summaries. Additionally, mechanisms for accountability should be established to address errors or unintended consequences and to enable recourse for affected individuals.

4. Plagiarism and Attribution:

The use of summarization models may raise concerns related to plagiarism and proper attribution. When users rely on summaries generated by the model, it is essential to emphasize the importance of giving appropriate credit to the original authors and sources. Plagiarism detection tools and guidelines for ethical content usage can help mitigate these concerns.

5. User Understanding and Informed Consent:

Users of the summarization model should have a clear understanding of how their data is processed and used. Providing user-friendly explanations and obtaining informed consent for data processing are ethical imperatives. Users should be aware of the model's capabilities, limitations, and potential impacts on their information consumption.

6. Accountability for Misuse:

While the model has numerous legitimate applications, it can also be misused for unethical purposes, such as spreading misinformation or generating misleading summaries. Developers and users should be aware of the ethical responsibility to use the model responsibly and refrain from activities that could harm individuals or society.

7. Accessibility and Inclusivity:

Ensuring that the benefits of the summarization model are accessible to all individuals, including those with disabilities or language barriers, is an ethical consideration. Efforts should be made to design interfaces and outputs that are inclusive and considerate of diverse user needs.

In conclusion, addressing the ethical aspects of the extractive text summarization model is integral to its responsible development and deployment. By actively considering issues related to bias, privacy, transparency, plagiarism, user understanding, accountability, and accessibility, stakeholders can work together to harness the model's potential for societal benefit while minimizing ethical risks and challenges.

5.4 Sustainability Plan

A sustainability plan is essential to ensure that the extractive text summarization model's positive impacts on society and the environment are maintained over the long term. Sustainability encompasses ethical, social, and environmental dimensions, and a comprehensive plan addresses these aspects.

1. Ethical Sustainability:

Bias Mitigation: Continuously monitor and update the model to detect and mitigate biases. Implement fairness-aware algorithms and bias reduction techniques to ensure that summaries are impartial and inclusive.

Transparency Measures: Maintain transparency in the model's processes and decision-making. Regularly communicate updates, improvements, and any ethical guidelines to users to foster trust and accountability.

Ethical Guidelines: Develop and enforce clear ethical guidelines for users, emphasizing responsible usage, proper attribution, and respect for copyright and privacy.

2. Environmental Sustainability:

Energy Efficiency: Continue to optimize the model's algorithms and infrastructure for energy efficiency. Explore renewable energy sources for data centers and computational processes to minimize the model's carbon footprint.

Electronic Waste Management: Promote responsible e-waste management practices among users. Encourage recycling and proper disposal of electronic devices to offset any potential increase in electronic waste.

Paper Reduction Initiatives: Collaborate with organizations to promote paper reduction initiatives. Encourage the adoption of digital summaries as an eco-friendly alternative to printed documents.

3. Social Sustainability:

Accessibility: Ensure that the summarization model remains accessible to a wide range of users, including those with disabilities or language barriers. Continuously improve user interfaces and accessibility features.

User Education: Develop educational materials and resources to help users make the most of the summarization model. Provide guidance on responsible content usage, citations, and ethical considerations.

Community Engagement: Foster a community of responsible users and developers who actively contribute to the model's sustainability and ethical usage. Encourage collaboration and feedback mechanisms.

4. Economic Sustainability:

Funding and Resources: Secure long-term funding and resources to support ongoing research, development, and maintenance of the model. Diversify funding sources to reduce dependency on a single entity.

Business Models: Explore sustainable business models that align with the model's ethical and environmental goals. Consider partnerships with organizations that share similar values.

5. Continuous Improvement:

Research and Development: Invest in ongoing research and development to enhance the model's capabilities, including bias detection and reduction, privacy protection, and energy efficiency.

Feedback Loops: Establish feedback loops with users and stakeholders to gather input and insights for continuous improvement. Actively seek user feedback to address emerging ethical and environmental challenges.

Collaboration: Collaborate with academic institutions, research organizations, and industry partners to stay at the forefront of technology and ethics in natural language processing and text summarization.

By implementing a sustainability plan that addresses ethical, environmental, social, and economic dimensions, the extractive text summarization model can contribute to a positive and lasting impact on society and the environment. Sustainability ensures that the model's benefits are realized while mitigating potential risks and challenges.

CHAPTER 6:

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

This section provides a comprehensive summary of our study on extractive text summarization, highlighting the key objectives, methodologies, findings, and contributions of our research.

Objectives and Scope:

The primary objective of our study was to develop and evaluate an extractive summarization model tailored for English news articles. Our research aimed to address the growing challenge of information overload in the digital age by creating a tool capable of condensing vast amounts of textual information into concise and coherent summaries. We focused on the following key aspects:

Implementation and Evaluation: We sought to implement and evaluate the effectiveness of the TextRank algorithm and Cosine Similarity measures in the context of summarizing English news articles.

Comparative Analysis: We conducted a comparative analysis of our model's performance against manually generated summaries and a benchmark built-in summarizer, Summy, to assess its accuracy, coherence, and conciseness.

Category-Wise Analysis: We explored how our model performed across different news categories, providing insights into its adaptability and robustness.

Methodologies and Approaches:

To achieve our objectives, we employed the following methodologies and approaches:

Data Collection: We collected a comprehensive dataset of English news articles from various online platforms, spanning categories such as business, entertainment, politics, sports, and technology.

Data Preprocessing: We performed essential data preprocessing steps, including spell correction, removal of special characters, punctuation, and text normalization. Unlike traditional NLP tasks, we retained stop words in our analysis due to their importance in news summarization.

Extractive Summarization: We applied the TextRank algorithm and Cosine Similarity measures to rank and select sentences for summarization. This process involved vectorization, cosine similarity computation, and the creation of a similarity matrix.

Findings and Insights:

Our study yielded several significant findings and insights:

The TextRank algorithm, in conjunction with Cosine Similarity measures, demonstrated promise in generating extractive summaries of English news articles.

Our model's performance, particularly in cases where the similarity score exceeded 0.5, indicated its potential to produce coherent, concise, and accurate summaries.

Category-wise analysis revealed variations in performance across different news genres, highlighting the model's adaptability to diverse content types.

Comparative evaluation against Summy, a benchmark summarizer, provided insights into our model's competitive performance and potential for further refinement.

Contributions:

Our research contributes to the field of extractive text summarization in the following ways:

We provide a practical and effective approach for summarizing English news articles, addressing the challenges of information overload and content diversity.

We emphasize the importance of ethical considerations in text summarization, advocating for responsible usage, transparency, and bias mitigation.

Our study offers a foundation for future research in areas such as bias detection and mitigation, multilingual summarization, user customization, visual summarization, and real-time summarization.

In summary, our study advances the field of extractive text summarization by presenting an effective model for condensing English news articles. By addressing key research objectives, methodologies, findings, and contributions, our study offers valuable insights and directions for further advancements in this domain.

6.2 Conclusions

In this section, we draw overarching conclusions from our research on extractive text summarization and summarize the key findings and implications of our study.

Effectiveness of TextRank Algorithm and Cosine Similarity Measures:

Our research focused on the application of the TextRank algorithm and Cosine Similarity measures in the context of extractive text summarization. We found that these techniques demonstrated promise in generating concise and coherent summaries of English news articles. The TextRank algorithm's ability to rank sentences based on their importance, coupled with Cosine Similarity measures for assessing sentence similarity, proved effective in identifying key sentences for summarization.

Performance Across Different Categories:

We conducted a category-wise analysis to evaluate our model's performance across various news genres, including business, entertainment, politics, sports, and technology. The

results revealed that the model exhibited adaptability, with varying levels of performance across different categories. This adaptability is a valuable characteristic for a summarization tool, as it enables the effective summarization of diverse content types.

Comparative Analysis with Summy:

Our study included a comparative analysis with Summy, a well-established built-in summarizer. This comparison provided insights into the competitive performance of our model in terms of accuracy, coherence, and conciseness. We found that our model performed competitively against Summy, underscoring its potential as an effective summarization tool.

Ethical Considerations:

Throughout our research, we emphasized the importance of ethical considerations in text summarization. We advocated for responsible usage of summarization models, transparency in processes, and the need to address biases in the summarization outputs. These ethical aspects are essential for ensuring that summarization technology benefits society without perpetuating harmful biases or misinformation.

Recommendations for Future Research:

Our study opens avenues for future research in the field of extractive text summarization. We recommend exploring advanced techniques for bias detection and mitigation, expanding the model's capabilities to handle multilingual content, allowing user customization of summarization preferences, integrating visual elements into summaries, and developing real-time summarization capabilities.

Implications for the Field:

Our research contributes to the field of extractive text summarization by providing a practical and effective approach to summarizing English news articles. We highlight the importance of ethical considerations, transparency, and responsible usage in the development and deployment of summarization models. Our study offers valuable insights and directions for further advancements in this domain.

In conclusion, our research demonstrates the effectiveness of the TextRank algorithm and Cosine Similarity measures in extractive text summarization. We emphasize the adaptability of our model across different content categories and underscore the significance of ethical considerations in the field. By drawing conclusions based on our findings, we pave the way for responsible and innovative advancements in text summarization technology.

6.3 Implication for Further Study

Our research on extractive text summarization has uncovered valuable insights and avenues for further exploration. In this section, we outline the implications of our study for future research in this domain, highlighting key areas that warrant additional investigation.

1. Bias Detection and Mitigation:

One critical implication for further study is the need to delve deeper into bias detection and mitigation techniques in extractive text summarization. While our research acknowledged the importance of ethical considerations, there is a growing demand for models that can not only identify biases in source texts but also rectify them in the summarization output. Future research can explore advanced algorithms and methodologies for bias detection and mitigation, with a focus on ensuring fairness, impartiality, and accuracy in summaries.

2. Multilingual Summarization:

Our study primarily focused on English-language news articles. However, the digital landscape is multilingual, and there is a pressing need for summarization models capable of handling content in multiple languages. Future research can extend our approach to develop multilingual summarization models, catering to a global audience and addressing the information needs of diverse language communities.

3. User Customization:

As users have varying reading habits and content preferences, there is potential for research in the area of user-customizable summarization. Future studies can explore techniques that allow users to tailor summarization outputs to their specific requirements. This may involve adjusting the length, style, or depth of summaries, providing a more personalized reading experience.

4. Visual Summarization:

While our study focused on text-based summarization, the integration of visual elements such as images, charts, and graphs into summaries can enhance content understanding. Future research can investigate methods for combining text and visuals in summarization, providing users with more comprehensive and informative summaries, especially for content rich in visual data.

5. Real-time Summarization:

The digital news landscape is dynamic and ever-evolving. Future research can explore the development of real-time summarization capabilities that provide up-to-the-minute summaries of rapidly unfolding news stories. Real-time summarization presents unique challenges and opportunities, making it an exciting area for further study.

6. Enhanced Ethical Frameworks:

As the ethical implications of AI and NLP technologies continue to be a focal point, future research can contribute by developing enhanced ethical frameworks for text summarization models. These frameworks should address not only bias and fairness but also issues related to transparency, accountability, and responsible usage. Research in this area can help guide the responsible development and deployment of summarization technology.

7. Collaboration and Interdisciplinary Research:

The field of extractive text summarization benefits from collaboration between researchers, industry partners, and user communities. Future studies can explore interdisciplinary approaches that incorporate insights from linguistics, psychology, journalism, and other fields to enhance summarization models. Collaborative research can lead to more holistic and effective summarization solutions.

8. Robustness and Adversarial Testing:

To ensure the reliability and security of summarization models, future research can focus on robustness testing and adversarial evaluation. This includes assessing how models perform under adversarial conditions and developing mechanisms to detect and mitigate potential vulnerabilities, such as adversarial attacks on summarization outputs.

In conclusion, our study on extractive text summarization offers a foundation for future research in several key areas. By exploring bias detection and mitigation, multilingual capabilities, user customization, visual summarization, real-time summarization, enhanced ethical frameworks, collaboration, and robustness testing, researchers can contribute to the continued advancement of text summarization technology and its responsible and effective use in the digital age.

APPENDIX

In this section, we provide supplementary information and materials to enhance the understanding and completeness of this research report. The appendices include additional details, data, and resources related to the study, supporting the findings and conclusions presented in the main body of the report.

Appendix A: Data Collection Sources

Appendix A lists the sources of data used for this research, including the online platforms from which news articles were collected. It provides information on the diversity of sources and the range of topics covered, contributing to the comprehensiveness of the dataset.

Appendix B: Data Preprocessing Details

Appendix B offers a comprehensive overview of the data preprocessing steps applied to the collected news articles. It includes detailed descriptions of spell correction, special character removal, text normalization, and other preprocessing techniques employed to prepare the data for analysis.

Appendix C: Summarization Model Details

Appendix C provides in-depth insights into the architecture and workings of the extractive text summarization model developed for this research. It includes details on the TextRank algorithm, Cosine Similarity measures, and the calculation of similarity scores.

Appendix D: Experimental Setup and Parameters

Appendix D outlines the experimental setup, including the hardware and software configurations used for conducting the experiments. It also specifies the parameters and settings applied in the summarization model during testing.

Appendix E: Raw Experimental Data

Appendix E presents the raw experimental data, including similarity scores, for each news article used in the study. This data serves as a reference for further analysis and validation.

Appendix F: Ethical Considerations

Appendix F addresses the ethical considerations and guidelines adhered to during the course of this research. It includes details on data privacy, informed consent, and other ethical aspects of the study.

Appendix G: Sustainability Plan Details

Appendix G provides a comprehensive sustainability plan, outlining measures taken to minimize the environmental impact of this research. It includes details on paper and energy conservation, recycling, and responsible resource usage.

Appendix H: Survey Questionnaires

Appendix H includes copies of any survey questionnaires or feedback forms used to collect data or opinions from human evaluators or participants during the research.

Appendix I: Supplementary Figures and Visuals

Appendix I contains supplementary figures, graphs, and visuals that complement the visual representation of data and analysis presented in the main body of the report.

These appendices collectively enrich the research report by offering transparency, additional context, and access to valuable supporting materials for readers and researchers interested in further exploring the study's findings and methodologies.

@Daffodil International University

REFERENCES

1. N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), January 2017, pp. 1-6.
2. R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D. Cavalcanti, et al., "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5755-5764, 2013.
3. V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, 2010.
4. J. N. Madhuri and R. G. Kumar, "Extractive text summarization using sentence ranking," in 2019 International Conference on Data Science and Communication (IconDSC), March 2019, pp. 1-3.
5. S. R. Rahimi, A. T. Mozhdehi, and M. Abdolahi, "An overview on extractive text summarization," in 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), December 2017, pp. 54-62.
6. A. M. Abu Nada, A. Alsaqqa, E. Alajrami, and S. S. Abu-Naser, "Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach," *Conference Paper*, September 2020.
7. B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," *Information Processing & Management*, vol. 57, no. 6, p. 102359, 2020.
8. M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, pp. 1-66, 2017.
9. C. Fang, D. Mu, Z. Deng, and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Systems with Applications*, vol. 72, pp. 189-195, 2017.
10. W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.
11. B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multi-document extractive text summarization: A comparative assessment on features," *Knowledge-Based Systems*, vol. 183, p. 104848, 2019.

12. A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders," *Expert Systems with Applications*, vol. 129, pp. 200-215, 2019.
 13. R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7764-7772, 2009.
 14. A. Jain, D. Bhatia, and M. K. Thakur, "Extractive text summarization using word vector embedding," in *2017 International Conference on Machine Learning and Data Science (MLDS)*, December 2017, pp. 51-55.
 15. N. S. Shirwandkar and S. Kulkarni, "Extractive text summarization using deep learning," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, August 2018, pp. 1-5.
 16. D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda, "Extractive text summarization system to aid data extraction from full text in systematic review development," *Journal of Biomedical Informatics*, vol. 64, pp. 265-272, 2016.
 17. Y. K. Meena and D. Gopalani, "Evolutionary algorithms for extractive automatic text summarization," *Procedia Computer Science*, vol. 48, pp. 244-249, 2015.
 18. A. Mahajani, V. Pandya, I. Maria, and D. Sharma, "A comprehensive survey on extractive and abstractive techniques for text summarization," in *Ambient Communications and Computer Systems: RACCCS-2018*, 2019, pp. 339-351.
 19. N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, March 2016, pp. 1-7.
 20. M. Cao and H. Zhuge, "Grouping sentences as better language unit for extractive text summarization," *Future Generation Computer Systems*, vol. 109, pp. 331-359, 2020.
- Here are the citations you provided converted into IEEE format:
21. P. Verma, S. Pal, and H. Om, "A comparative analysis on Hindi and English extractive text summarization," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, no. 3, pp. 1-39, 2019.

22. W. M. Wang, Z. Li, J. W. Wang, and Z. H. Zheng, "How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds," *Expert Systems with Applications*, vol. 90, pp. 439-463, 2017.
23. S. A. Babar and P. D. Patil, "Improving performance of text summarization," *Procedia Computer Science*, vol. 46, pp. 354-363, 2015.
24. A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, and A. Basit, "Extractive text summarization models for Urdu language," *Information Processing & Management*, vol. 57, no. 6, p. 102383, 2020.
25. S. S. Naik and M. N. Gaonkar, "Extractive text summarization by feature-based sentence extraction using rule-based concept," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2017.
26. R. Rani and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," *Expert Systems with Applications*, vol. 186, p. 115867, 2021.
27. Y. Ledeneva, A. Gelbukh, and R. A. García-Hernández, "Terms Derived from Frequent Sequences for Extractive Text Summarization," *Lecture Notes in Computer Science*, pp. 593–604, [n.d.].
28. N. Chatterjee, A. Mittal, and S. Goyal, "Single document extractive text summarization using Genetic Algorithms," in *2012 Third International Conference on Emerging Applications of Information Technology*, 2012.
29. A. R. Mishra, V. Panchal, and P. Kumar, "Extractive Text Summarization - An effective approach to extract information from Text," in *2019 International Conference on Contemporary Computing and Informatics (IC3I)*, 2019.
30. D. Suleiman and A. A. Awajan, "Deep Learning Based Extractive Text Summarization: Approaches, Datasets and Evaluation Measures," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019.

EXTRACTIVE TEXTRANK-BASED NLP NEWS SUMMARIZATION FOR MULTIPLE DOMAINS

ORIGINALITY REPORT

19% SIMILARITY INDEX	15% INTERNET SOURCES	10% PUBLICATIONS	12% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
3	Bilal Khan, Zohaib Ali Shah, Muhammad Usman, Inayat Khan, Badam Niazi. "Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey", IEEE Access, 2023 Publication	1%
4	www.irjmets.com Internet Source	<1%
5	Submitted to University of Sydney Student Paper	<1%
6	www.researchgate.net Internet Source	<1%
7	Milad Moradi, Maedeh Dashti, Matthias Samwald. "Summarization of biomedical articles using domain-specific word	<1%