ASSESSMENT OF EVALUATION METRICS USED FOR ABSTRACTIVE TEXT SUMMARIZATION AND A PROPOSAL OF NOVEL METRIC

by

Priyanka Ramane

DISSERTATION

Presented to the prof

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

November 2024

ASSESSMENT OF EVALUATION METRICS USED FOR ABSTRACTIVE TEXT SUMMARIZATION AND A PROPOSAL OF NOVEL METRIC

by

Priyanka Ramane

APPROVED BY

Milica Popović Stijačić, PhD, Chair

RECEIVED/APPROVED

BY: XXXX PhD

Chair

DEDICATION

To God, whose guidance and blessings have led me through every step; to my family, whose unwavering support made this journey possible; and to my sister, whose encouragement and love have been my constant source of strength.

ACKNOWLEDGMENTS

First and foremost, I express my deepest gratitude to God for the unending love, guidance, and blessings throughout this journey. The strength and wisdom have been my cornerstone, giving me the perseverance to overcome the challenges during my doctoral studies.

To my family, I am eternally grateful for your unwavering support and encouragement. To my parents, thank you for instilling in me the values of hard work and determination. Your sacrifices and belief in my abilities have been a tremendous source of motivation. To my sister, your kindness and patience have been invaluable. Your support has meant the world to me.

I extend my heartfelt thanks to SSBM University for the opportunity to pursue my doctoral studies. I am grateful to all the faculty and staff members who have contributed to my academic growth.

A special note of gratitude goes to my supervisor, Professor Mario Silic, for his invaluable guidance, mentorship, and support throughout this journey. Thank you for believing in me and pushing me to achieve my best.

ASSESSMENT OF EVALUATION METRICS USED FOR ABSTRACTIVE TEXT SUMMARIZATION AND A PROPOSAL OF NOVEL METRIC

ABSTRACT

Background

Automatic text summarization is one of the most efficient techniques to obtain a shorter and a more concise version of the unstructured text data. Text summarization algorithms have seen a huge recent increase in performance following investments in very advanced Machine Learning, Large Language Models and Generative AI. Even though these text summarization algorithms are getting improved, the way we measure the accuracy of text summary generated by these algorithms remains unchanged. We still use a traditional method that mostly uses the word overlaps in generated summary and original summary. There is a constraint in analyzing the summary produced. This paper examines these measures in the business arena which are unable to serve business needs like relevance of summary, coherence, and informativeness. As businesses leverage transformer-based models to generate customer feedback and insights, there arises a need for improved evaluation metrics. This research suggests a new metric that matches business objectives and enhances automated summarization to gain a competitive edge.

Methods

This research study uses quantitative methods to evaluate a metric and improve text summarization evaluation. The study was conducted based on the traditional metrics like ROUGE and BLEU and proposed a new metric called the Unified Summary Evaluation Score (USES). The study examined BART and T5 model who are capable of generating text summaries from large datasets. Statistical techniques (t-tests, ANOVA, etc) were used to

assess the performance and consistency of USES against other conventional metrics, with respect to both surface-level accuracy and semantic-level accuracy.

Results

USES performance metrics yield better semantic accuracy and more reliable summaries than traditional metrics, according to the analysis of quantitative data. The combination of BERTScore, Wu-Palmer similarity and cosine similarity enabled USES to provide a more holistic evaluation along with better accuracy and consistency. According to statistical tests, USES has greater consistency in evaluating abstractive summaries and aligns better with humans.

Discussion and Conclusion

According to the results, USES provides a more complete evaluation of summaries which is better than the traditional metrics that only look at the overlapping words. Through integration of semantics, coherence as well as coverage, USES provides a more accurate and fair assessment of summaries generated by various large language models. This study helps develop better assessment frameworks in terms of text summarization which will be useful for various natural language processing and machine learning applications. The study establishes a groundwork for further investigations into evaluation strategies that are efficient and scalable to improve summarization.

Keywords

Automatic Text Summarization, Natural Language Processing (NLP), Unstructured Text Data, Text Summarization Algorithms, Machine Learning, Large Language Models, Generative AI, Evaluation Metrics, Word Overlap Methods, Transformer-based Models.

LIST OF ABBREVIATIONS

NLP: Natural Language Processing

GAN: Generative Adversarial Network

BERT: Bidirectional Encoder Representations from Transformers

GPT: Generative Pre-trained Transformer

BERT: Bidirectional Encoder Representations from Transformers

T5: Text-To-Text Transfer Transformer

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

WUP Similarity: Wu-Palmer Similarity

BLEU: Bilingual Evaluation Understudy

List of Tables

List of Figures

TABLE OF CONTENTS

Chapter 1 INTRODUCTION

1.1 Introduction

In today's digital era, data is said to be the new fuel (Humby, 2006). Earlier, data analysis was possible only for structured data however with the surge of machine learning, deep learning and natural language processing; analysis over unstructured data which is in the form of text, images and videos is also possible.

Progress has been made in natural language-related disciplines, spanning from natural language processing to natural language understanding and natural language generation (Howard and Ruder 2018; Vinyals and Le 2015; Wang et al., 2019). There is a continuous progress towards multiple successful research on improving performance of various NLP techniques, however, these newly research based techniques are evaluated on few industry standard evaluation metrics. Single NLP task can be achieved via multiple techniques however the efficacy of the task is evaluated based on few fixed sets of metrics only. Text summarization is one such technique that can be achieved via traditional rule based lexical morphological methods, or involving various semantic-based methods to capture the meaning or syntactic techniques to capture the grammar rules of the languages. In the most recent days, seq2seq models which utilize recurrent neural networks for sequence generation tasks, transformer-based language models which uses self-attention mechanisms; capture the various aspects of the languages to perform text summarization. However, to evaluate the model performance, generated summary is still evaluated on few intrinsic techniques such as BLEU (Lin, 2004) and ROUGE (Papineni et al., 2002) Score. BLEU and ROUGE score heavily depend on the occurrences of the words in the generated text output, comparing them with reference summary to calculate the accuracy. These evaluation metrics often fail to capture the lexical and semantic aspect of the generated text.

1.2 Research Problem

In recent years, automatic text summarization has taken a big leap from traditional rule based methods to deep learning and LLMs. Even after witnessing these advancements, the assessment of generated summaries is still done using traditional metrics. Most of the times BLEU and ROUGE are used to measure the similarity score between the reference and the generated summary. They mainly measure the visual word overlap between the words used in the predicted summary and the words used in the ground-truth summary. These metrics have shortcomings in evaluating the quality of abstractive summaries that look at the meaning rather than the words. The lack of evaluation methods results in bias and does not give the correct assessment of summaries, especially in reference to their lexical and semantic quality. In addition, existing metrics cannot fully capture the linguistic subtleties and variety of effective summarization. They are not always complete and therefore confound a model's assessment. There is a strong need for smarter evaluation metrics that will provide a more comprehensive understanding of summary quality, accuracy, consistency and thus, summary models can be built better.

1.3 Purpose of Research

This paper aims to demonstrate and propose a novel evaluation of text summarization technique which can overcome the shortcomings of present-day techniques like BLUE and ROUGE. A new approach to evaluate summaries generated by large language models that takes into account the lexical and semantic aspects. The research aims to improve the general quality and trustworthiness of summarization models by enhancing evaluation methods. In the end, this paper connects the advanced models for summarization with their corresponding evaluation metrics.

1.4 Specific Aims

- **Check the shortcoming of existing metrics:**
    - Look at how the traditional metrics BLEU and ROUGE do not fully capture different aspects of the summaries generated by LLMs.
    - Learn how they affect the evaluation of the lexical and semantic quality of abstractive summaries.
- **Develop an Evaluation Metrics:**
    - A new assessment metric must evaluate the semantic quality and coherence quality in a summary evaluation.
    - Make sure these metrics consider lexical diversity and accurate, consistent, providing a more holistic view of summary.
- **Check and compare new metrics:**
    - Test and validate the evaluation metrics through experimental studies.
    - Check how these new metrics do against the old ones to show that they are much better.
- **Guidelines for Metric Selection:**
    - Create guidelines to select best evaluation metrics for different types of summarization experiments.

1.5 Significance of the Study

Through this study, we aim to contribute to automatic text summarization by designing new evaluation metrics.

- **Enhanced Evaluation Accuracy:** New metrics that capture aspects of lexical and semantic will be developed in this research to get a more accurate assessment of the quality of summary. This helps researchers to understand strengths and weaknesses of their summarization models.

- **Improved Summarization Models:** More sophisticated and thorough evaluation metrics will accelerate better models on summarization. Accurate evaluation metrics can provide precise indications of where the model works well and where it does not, guiding subsequent refinement of the model.
- **Broader Applicability:** The metrics we propose can be used for all kinds of summarization tasks whether it's an extractive one or an abstractive one. Their versatility will make them useful for a variety of purposes including but not limited to educational and business applications.
- **Reduction of Bias:** Metrics like BLEU, ROUGE measure translation and summarization quality based on word overlap. This may introduce intrinsic bias. The new metrics will aim to reduce such biases by taking into account the meaning and coherence of the generated summaries that will lead to fair evaluations.
- **Advancement of NLP Research:** This research will advance the field of natural language processing (NLP) by establishing new benchmarks for evaluation. Better metrics will help not only summarization but also other Natural Language Processing (NLP) tasks where semantic and lexical accuracy are concerned.
- **Guidance for Future Research:** This research will create a framework for future research by developing principles for identifying optimal evaluation metrics. This work will help researchers to gain more clarity as to which metrics to use for which types of summarization experiments.
- **Practical Impact:** The findings of this research may assist in the application of any real-life cases where accurate text summarization is required for news aggregation, literature reviews, and legal summarization. With better evaluation metrics, the quality and usability of the summaries will be improved, improving the end-user experience.

1.6 Research Purpose and Question/Hypothesis

The study will particularly focus on this research gap and problem statement to propose a novel method for text summarization evaluation metric that incorporates the lexical and semantic aspects of the summary. Below are the research questions that will be addressed during this study.

1. How can we evaluate summaries produced by large language models using different metrics or evaluation strategies not limited to BLEU and ROUGE?
2. What other evaluation techniques can look at the meaning of summaries (semantic) and their coherence beyond simple word overlap measurements?
3. What factors should guide the selection of a set of optimal text summarization evaluation metrics in various summarization experiments?

Chapter 2 LITERATURE REVIEW

2.1 Introduction

In the era of Artificial intelligence, analysis over the unstructured data such as text, images, speech and videos are possible. Many recent advancements in the field of natural language processing in conjunction with Artificial intelligence focuses on improving various task such as text generation, text summarization, text classification and so forth. Generative AI is one of the most groundbreaking advancements of the era which uses transformer based LLMs (Large Language models) which generate new context on human instructions. Though there are many developments in content generation, there is a notable lack of substantial initiatives aimed at developing novel evaluation metrics capable of assessing the performance of language models like LLMs and the content they generate. While current LLMs demonstrate proficiency in producing text, speech, images and videos; our research concentrates on LLMs specifically engaged in text summarization (text generation). We aim to assess the appropriateness of existing state-of-the-art summarization evaluation metrics. These metrics have limitations primarily relying on traditional methods for accuracy assessment. Given the transformative impact of Generative AI on summary generation, it is imperative to adapt text summarization evaluation metrics to align with these advancements.

The chapter “Literature Review” started with the introduction of “Text Summarization” and different type of summarization and its usage in industry. In next phase of this chapter, state of art algorithms in AI, Gen AI and LLMs are discussed along with contributions to text Summarization. This phase also covers how over the period text summarization algorithms have evolved. It also includes the limitations and challenges of LLM for text summarization. The understanding of Text Summarization using LLMs will lay foundation for next phase of literature review which will touch base on the various evaluation metrics that are currently being used. This phase of the study will cover history of various text summarization metrics,

their efficacies, accuracies and boundaries. Furthermore, detailed literature review is carried out to identify the research gap between the LLM generated summary and current state of the art text summarization evaluation metrics. Here onwards, based on these research gaps, study explains the necessity of new text summarization evaluation metrics that will overcome the limitation of existing evaluation metrics and LLM.

2.2 Theoretical Literature Review

This chapter is critical analysis of existing research and scholarly works relevant to text summarization and current state of text summarization evaluation metrics. The goal is to present a coherent and well-organized summary of the literature.

2.2.1 Text Summarization

In today' digital era, we are surrounded by many digital equipment, social platforms which generate vast amount of unstructured text data. Digital newspaper, contract documents, various study documents, historic monuments are available in text format, however, it is nearly impossible to understand the gist of these documents, social media data without actually reading them. "Text Summarization" helps creating the summarized and meaningful compact version of these text which could be very well understood and further processed for various downstream work. Loosely, text summarization can be defined as shorting the long text into fluent summary. Though traditionally text summarization used to be performed with manual efforts; nowadays the same work is done via Automatic text summarization. "Automatic text summarization" technique can be defined as the machine generated meaningful shortened summary of the long text. Automatic text summarization retrieves and generates the summary in short span with almost zero human intervention. Automatic text summarization can be achieved by Rule based algorithms, Machine learning based algorithms or with state of the art Transformed based LLMs (Generative AI) algorithms. In the next section, we will study the various types of Automatic Text summarization.

2.2.2 Type of Automatic Text Summarization

Types of Summarizations Based on Summary Creation:

- **Abstractive Summarization:** Abstractive Summarization is a method of generating a summary from the given text. It comes under the domain of Natural Language Processing. The summaries created through this method include new words which are not part of the input document and have a similar meaning. Though the generated sentences are considered grammatical theoretically, practical limitations of text generation techniques may require further enhancement to align with human-generated summary.
- **Extractive Summarization:** Extractive summarization is used mainly when the text data is unstructured. The strategy of extractive summarization is to retrieve the text constituents (phrases, keywords, etc.) from the entirety of the text to summarize the document. The significance of these components is usually established by means of rules based on grammar and lexical knowledge.

Types of Summarizations Based on Document Type:

- It takes input in one language and gives output in the same language. It only focuses on one language. This kind of summarization is **Mono language Summarization**.
- This type of multi-lingual summarization refers to developing a unified system which allows the summarization algorithm to be applied to texts in several languages. In this case, the input text document and output summary are in the same language. This kind of summarization is **Multi language Summarization**.
- Aims to create summaries that are in different languages this implies that in such a case the input text and the output summary will be in two different languages. This kind of summarization is **Cross language Summarization**.

Types of Summarizations Based on Document Size:

- **Single Document Summarization:** Single Document Summarization is the simplest type of summarization that limits itself to one document. This can be one paragraph or one blog.
- **Multi Document Summarization:** The process of extracting one or several summaries that are non-redundant and provide an overview of text documents from multiple documents.

2.3 Application of text summarization

There are many useful applications we can see in industry for text summarization. Starting with simple application in the media field such as news summarization, social media blog summarizations. these will help end use to get shorten version of the information instead of relying on lengthy text. Generally, any newspaper is of 10 pages long however nowadays newer business takes this data and create shorter and crisp summary generating the same information in as small as 50 words news.

Text summarization also has great potential in education field. Here, it can help to create shorter version of the data such as textbook text, lectures, article and so on. In the field of legal, it does have many of the use-cases that revolves around contract summarization that helps lawyers, legal professional to understand the legality of the long documents.

Same goes in the field of healthcare where summarization helps to generate summary of patient records, medical research papers that can help medical profession to reach a decision-making point. This kind of summarization can be useful in healthcare research industry.

In customer support, customer review is the key to business success. however, going through all reviews may not be feasible. Also support tickets raised by the end use can also be key indicator of the business health however the volume of the data may not allow to go over

individual ticket. Summarization is the key in customer support where it helps to summarize all the feedback, tickets and reviews to create holistic point of view.

In general, as well for day-to-day activity like, generating meeting transcript would be very useful that can help user to find the points discussed, key decision taken and action needed. earlier all these actions were performed manually however, now through auto text summarization algorithms makes it easy to get this data to end user at click of button saving hours.

All these applications mentioned above are real lie examples where summarization is already exists and saving hours of the manual efforts. With increased precision and latest development coming in every day in the generative ai; text summarization will also go in the area like financial data review, governance policies and compliance data where data precision is key requirement.

2.4 Empirical Literature Review

Broadly there are two types for Abstractive summarization i.e. Rule Based and Machine Learning Algorithm based.

2.4.1 Rule Base Abstractive Summarization:

Rule-based summarization creates summaries for textual data in abstractive form through predefined rules. In Rule-based Abstractive Summarization, rule-based methods are used to guide or constrain the process of generating the abstractive summary. These rules may specify how to paraphrase/rewrite a sentence, what information takes precedence, how to treat certain linguistic units, etc. The aim is to combine the advantages of rule-based systems (clear guidelines, explicit knowledge) with the flexibility and creativity of abstractive summarization. Often Rule based summarizer are extractive in nature where handcrafted rules are identified to understand which sentence could be included as a part of final summary.

According to Mehdad et al., (2014), abstractive summarization can be achieved by using phrasal queries of Spoken and Written Conversations. The approach mentioned here focuses on novel abstractive query-based summarization system that ranks and extracts conversation utterances, clustering them based on lexical similarity. The resulting summary combines user-defined phrasal queries with the conversation's overall content. In some scenario rule-based summarizer would be applicable only to certain domain as rules are identified manually by domain experts and could be applicable to specific subject and domain. In one research paper, researchers (Pimpalshende and Mahajan, 2016) talk about one of such rule-based summarizers which is specific to summarize historic document. Though this research is based on extractive summarization, features and technique mentioned in this work could also be used for abstractive summarization. Another innovative approach has been proposed in paper (Le and Le, 2013) that employs discourse rules, syntactic constraints and a word graph to generate abstractive summary with promising results. Some researchers (Vodolazova and Lloret, 2019) discussed the enhancement of abstractive text summarization through the integration of syntactic text simplification, subject-verb-object concept frequency scoring and rules for transforming text into its semantic representation. Another way to do text summarization using artificial intelligence based on natural language processing (NLP) was proposed (Yahya Saeed et al., 2021). This method takes in TF-IDF, PageRank keywords, a sentence score algorithm and Word2Vec word embedding. It addressed the limitations in providing the basic theme of the documents. The approach could generate keywords of varied lengths. This improved the similarity of metadata and attempted to solve the challenge of determining the representative keyword. The examination further showed that the proposed abstractive summarization using deep learning principles, had longer matches with other similar techniques and provided a consistent measure of similarity in comparison assessments. The paper (Oya et al., 2014) developed a system that could automatically produce an abstractive

summary of the meeting conversations, through a multi-sentence fusion technique that could produce abstract templates. Researchers looked at summary-source meeting transcript relationships to determine the most effective template used in the system. The results were successful in terms of readability and informativeness through manual and automatic evaluations. According to another paper for abstractive summarization (Kallimani, Srinivasa and Eswara Reddy, 2016), a document condensing process with key information extraction used a unified model with attribute-based Information Extraction (IE) rules and class templates. Using the TF/IDF rules for classification and the lexicon analysis for strong IE rules, it could handle the complexities of Indian languages.

2.5 Machine Learning Based Abstractive Summarization:

Machine Learning based Abstractive Summarization overcame the limitations of Rule based Summarization. Machine learning-based abstractive summarization refer to the use of machine learning algorithms and models to generate the summary with rephrasing of the original text. These algorithms are pre-dominantly based on neural network and deep learning that are capable of capturing contextual information of the word to generate abstracted and paraphrased summary. Many studies used Neural Attention Model to generate abstractive sentence summarization. It uses a local attention-based model (Rush, Chopra and Weston, 2015) o generate each word of the summary. One such work (Nallapati et al., 2016) talked about the use of abstractive text summarization using Attentional Encoder-Decoder Recurrent Neural Networks. In 2017, transformer was introduced in paper (Vaswani et al., 2017), a novel architecture solely based on attention mechanism outperformed machine translation task. This transformer-based architecture was further used as unit component in many large language models such as BERT (Devlin et al., 2018) and ChatGPT (ChatGPT, 2023). Another approach, BART (Lewis et al., 2019) introduces a denoising autoencoder pre-training objective for sequence-to-sequence tasks. It has been applied successfully to abstractive summarization

demonstrating the effectiveness of this pre-training approach in generating coherent and informative summaries. PEGASUS (Zhang et al., 2020) employs a gap-sentence generative pre-training approach to train a transformer-based model for abstractive summarization. By leveraging extractive pre-training, PEGASUS achieved state-of-the-art performance on various summarization benchmarks. T5 (Raffel et al., 2019) introduced a text-to-text framework where every NLP task including summarization is cast as a text generation problem. This flexible approach simplifies the training process and has been successfully applied to abstractive summarization tasks, showcasing the versatility of transformer-based models.

2.6 Evolution of Evaluation Metrics for Abstractive Summarization

As we have seen in earlier module, the landscape of the text summarization is very vast; researchers have tried generating the text summary by key phrase extraction, based on linguistic rules or finding morphological cue extraction. Also, recent days transformer-based language-based development focuses on learning the hidden patterns and handle word polysemy and word disambiguation to generate the best text summary. This implies that generated summary is dependent not only on lexical aspect but also on the semantic and syntactic aspect of the language. Though all above studied text summarization techniques consider all aspects of the language to include lexical, semantic and syntactic features to generate the summary; evaluation metrics used for these researches predominantly uses the word occurrences of expected summary and generated summary to measure the accuracy. Below are few frequently used Summarization Metrics:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures overlap of n-grams (sequences of words) between the generated summary and the reference summary. This metrics was proposed (Lin, 2004) to auto calculate the accuracy of a summary by comparing it to ideal human generated summary. These measures count

the number of overlapping units such as n-gram, word sequences and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

- **BLEU (Bilingual Evaluation Understudy):** Evaluates the precision of the generated summary by comparing it to one or more reference summaries. Pre-dominantly it is used for machine translation task however alternatively used to evaluate text summarization as well. Research mentioned in (Papineni et al., 2002) covers how BLEU score be applied for machine translation and text summarization. BLEU score is based on brevity penalty and the n-gram overlap, where the brevity penalty penalizes generated translations that are very short.
- **MoverScore** - MoverScore (Zhao et al., 2019) metrics considers combination of contextualized representations of system and reference texts and distance between these representations measuring the semantic distance between system outputs and references.
- **METEOR** – METEOR, (Banerjee and Lavie, 2005), is one of the earlier papers that proposed a method for evaluating machine translation. METEOR, which stands for "Evaluation of Translation with Explicit ORdering," is an automatic evaluation metric that is popular in Machine Translation. It evaluates automatic translations with respect to human translations (the reference translations).
- **BERTScore** – BERTScore (Zhang et al., 2020) calculates a similarity score for every token in the candidate sentence against every token in the reference sentence. It uses embeddings to determine similarity between tokens, unlike an exact match.

Above mentioned are most commonly used evaluation metrics for text summarization. However, evaluation metrics like ROUGH, BLEU, METEOR use word overlap method to calculate accuracy. They often generate the score based on counting the matching generated

words for evaluation and does not consider the syntactic and semantic aspects of the generated summary which is creating the evaluation bias. Summary generated by large language model are generalized over the vast corpus, resulting in generating new text as a summary. To evaluate this summary using only the count of matching words is major drawback and limitation.

On other hand, evaluation metrics like BERTScore, MoverScore uses embedding based semantic similarity method to calculate the accuracy. Since these are based on embeddings, they are sensitive to tokenization. Variations in tokenization approaches may impact the calculated scores.

2.7 Research Gap

While reviewing literature for text summarization techniques, a similar observation was made: the frequent use of modern machine learning techniques. At first, statistical or linguistic approaches were used for text summarization techniques. The methods usually consist of key-phrase extraction, high-frequency words identification, cue-phrase extraction from sentences to summarize the given document. However, with recent advancements in neural networks, various frameworks have been proposed to address text generation challenges. The emphasis in contemporary developments is on understanding hidden patterns, addressing word polysemy and handling word disambiguation, suggesting that generated summaries depend not only on lexical aspects but also on the semantic and syntactic aspects of language. However, the scarcity of research in text evaluation metrics has exposed a significant gap, presenting an opportunity for the development of innovative metrics which is discussed below in detail.

2.8 Evaluation Metrics Gap

Although summaries are generated after extensive consideration of lexical, semantic and syntactic aspects, most papers rely on ROUGE, BLEU only for evaluation. While reviewing

the literature, we analyzed summarization methods including rule-based, machine learning-based and novel transformer-based. Despite the method diversity, most research papers are assessed with normal ROUGE and BLEU score as evaluation metrics. The metrics mostly take into consideration the number of generated words that match and do not take into account the semantic and syntactic aspect of the summaries. As a result, there can be an evaluation bias that does not assist with proper evaluation of the summaries.

The advanced literature review indicates that the current state-of-the-art automatic text summarization is constrained by limitations in evaluation metrics. While some advanced metrics attempt to capture semantic distances, the prevailing evaluation criteria still fall short of encompassing all linguistic aspects of languages. Apart from linguistic standpoint, if we consider from a business perspective, these traditional metrics neglect critical elements such as alignment with organizational goals, decision-making utility and user engagement. For example, in applications like executive dashboards, customer service systems or social media analytics, summaries should bring actionable insights rather than simple text and words overlap. ROUGE and BLEU do not account for high stakes needs, such as ensuring consistency in legal or medical summaries or the dynamic requirements of businesses, like customized summaries for specific objectives. In brief, the evaluation metric should encompass the lexical and semantic elements of the summary; and that is currently absent.

This research aims to address this gap by thoroughly examining various metrics and recognizing the need for an improved text summarization evaluation metric. There is a significant potential for enhancing the functionality of existing evaluation metrics. The study will particularly focus on the latest transformer-based language models and propose a novel method for text summarization evaluation metric that incorporates the lexical and semantic aspects of the summary. This thesis seeks to contribute to a more comprehensive and accurate evaluation of text summarization techniques.

2.9 Conclusion

Natural Language Processing (NLP) is one of the streams or fields of Computer Science and Artificial Intelligence which manages unstructured text data. Because there is so much text data present; all the researchers are exploring ways to extract meaningful insights from the data. A primary challenge that NLP faces is the text generation. There are several applications such as text summarization, machine translation and question-answering systems which researchers have explored. Further, enterprises have also implemented them for automated downstream activities.

As machine learning progresses with the development of various algorithms, the field of Natural Language Processing evolves concurrently by incorporating these algorithms into text-related tasks. Despite these advancements, the evaluation metrics for newly developed text algorithms remain somewhat outdated, relying on conventional and traditional methods.

This study aims to examine metrics used in text summarization evaluation and their drawbacks in context with the modern-day text generation techniques. Over the years, text generation techniques in summarization have evolved and contain transformer-based language models. The evaluation metrics of text summarization are still largely dependent on word frequency counts to evaluate its performance.

This study examines the past evaluation techniques and highlights an effective metric for an experiment set-up. This study will also help to pick the right text summarization evaluation metrics for any experimentation.

Chapter 3 METHODOLOGY

3.1 Overview of Research Problem

The automatic text summarization of texts has changed from simple rule-based methods to more complex deep learning-based technologies, especially large language models (LLMs). With these LLM models, summarization has reached human-like intelligence to create comprehensive summaries for most of the documents. however, methods for evaluating summaries have not improved to same extent.

At the moment, the majority of summarization models, which use rules or LLMs, are evaluated using BLEU and ROUGE. The metrics work by looking at what words in the generated summary appear in the reference summary. This should work well for extractive summarization. These measures have limitations when dealing with abstractive summarization which mainly focuses on the perception of the text. The summary does not sufficiently assess how meaning deepens and how language flows, according to the writer. This issue is particularly pertinent for summaries produced by transformer-based models, which may incorporate words from outside the original text.

This research is proposing new metrics for evaluating LLM-generated summaries. This study will emphasize more on measuring meaning, understanding and overall coherence of summary rather than mere matching words. It will also assess what factors to consider for choosing the best evaluation methods for different summarization tasks ensuring the compatibility between the metrics and the advanced nature of new summarization models.

3.2 Operationalization of Theoretical Constructs

In this work, we improved the evaluation of summaries produced by LLMs. At the moment, the majority of evaluation measures focus on the number of overlaps between the created summary and original summary text. However, this does not necessarily imply that the summary reflects the essential meaning contained in this original text. To get the clear picture, we needed to practically define and measure some key concepts which are important.

Understanding the meaning of text was key to our evaluation and to check this, we needed to see whether the summary really captures the main ideas of the original text and not the actual words. Using some sophisticated tools, we compared the meanings of the summary and the source text. For example, using models like BERTScore that can tell us how similar the meaning of the summary is to the original. This helped us see if the summary accurately reflects the important points from the original text.

Coherence is all about how a summary is organized and how smoothly the ideas flow from one text to another. A coherent summary should be easy to follow. To measure coherence, we designed methods which could check whether sentences in summary connect well and whether the summary was logically ordered.

We also needed to assess the abstractive quality. The summary should not pick lines from the original material; instead, must rephrase the target material. To write a good abstraction summary, replacing the words and sentences may work but without changing the meaning. To measure this, we understood how well the summary used new words without missing the key information.

At last, this study checked if our evaluation metric aligns with the strengths of modern summarization models. This made sure that new metric was able to generate LLM generated summaries. Traditional metrics might not always recognize the advanced features of these

models. This study compared old metrics with new metric to see which ones better capture the quality and capabilities of LLMs in terms of understanding meaning and coherence.

3.3 Research Purpose and Questions

The thesis aims to introduce a new text summary evaluation framework which overcomes the limitation of existing evaluators like BLEU and ROUGE. This new way looked at many different things in the summaries, for example, word choice, meaning and sentence structure to better and more complete assessment of the summaries produced by large language models. Our goal was to improve evaluation metrics which in turn make models more reliable. In the end, we would like the evaluation metrics to also evolve to match the capabilities of summarization techniques.

3.4 Specific Aims

- **Check the shortcoming of existing metrics:**
    - Look at how the traditional metrics BLEU and ROUGE do not fully capture different aspects of the summaries generated by LLMs.
    - Learn how they affect the evaluation of the lexical and semantic quality of abstractive summaries.
- **Develop an Evaluation Metrics:**
    - A new assessment metric must evaluate the semantic quality and coherence quality in a summary evaluation.
    - Make sure these metrics consider lexical diversity and accurate, consistent, providing a more holistic view of summary.
- **Check and compare new metrics:**
    - Test and validate the evaluation metrics through experimental studies.

- Check how these new metrics do against the old ones to show that they are much better.

- **Guidelines for Metric Selection:**
  - Create guidelines to select best evaluation metrics for different types of summarization experiments.

3.5 Research Question

- What novel evaluation metrics or methods can be created for the shortcomings of BLEU and ROUGE with respect to evaluating large language model generated summaries? How can the newly developed metrics effectively capture the lexical and semantic information of the summaries?

- **Hypothesis**:

  Novel evaluation metrics will improve quality assessment of summaries produced by large language model (LLM) significantly against BLEU or ROUGE. The new metrics will measure not only lexical richness but also coherence and be consistent. The above metrics will better address the shortcomings of current methods by measuring aspects of quality such as meaning and coherence, rather than merely word overlap.

3.6 Research Design

Evaluation of any new metric is generally done using a Quantitative research design, as measuring the efficacy of new metrics requires statistical analysis to ensure objective results. In this thesis, a quantitative approach is proposed as the best way to evaluate new metrics for automatic text summarization. This method relies on measurable data to offer a clear and systematic evaluation of how well the new metrics assess the quality of summaries compared to traditional benchmarks.

3.6.1 Quantitative Research Design

The main objective of the quantitative approach was to measure the performance of the newly proposed metrics subjectively. These measures were evaluated against classical scores, for instance, BLEU, ROUGE, and so on across varied summarization systems to find out their robustness in capturing measures of summary accuracy, coherence and consistency.

- **Data Collection**: During this phase, the old as well as new metrics were used on already established datasets and summarization benchmarks like CNN/Daily Mail dataset (Hermann et al., 2015). BART and T5 were used to the test dataset to generate the new summary, and these summaries were collected and served as data input for our experiments to assess new metrics.
- **Analysis**: Initial data analysis was done to understand the input data very well and confirmed the relevance to be used in our experiment. These were done via visual tools such as boxplot and histogram. Later part in study, statistical approaches including t-test, ANOVA, other significance testing were used to compare performance of the new metrics with traditional ones. Statistical analyses were conducted to find out how strongly related the new measures were with human evaluations. Through these methods it was possible to assess how effective the novel metrics were in capturing other important features such as meaning, coherence, relevance and consistency.

With quantitative method, one can ensure that the new metrics evaluation was based on rigorous empirical evidence and objective measurement. This thorough assessment enhanced our understanding of how these new metrics perform in specific contexts, thus revealing their strengths and weaknesses. One could use this method to create trustworthy evaluation metrics that can enhance the quality assessment of summaries.

3.6.2 Population and Sample Selection

A quantitative research design was proposed for the evaluation of the new measures for automatic text summarization. The technical performance of the metrics measured through statistical techniques provides an objective and complete picture.

The quantitative design's population comprised of all text summaries generated by language model BART and T5 abstractive summarization model from any domain such as news articles, scientific papers, etc. The benchmark datasets consisting of CNN/Daily Mail dataset are well-known examples for data summarization.

The summaries produced by these models were evaluated using traditional metrics like BLEU and ROUGE, as well as by new metrics which was being developed. This quantitative technique allowed us to statistically check how well the new measures correlated against established ones in the ability to capture meaning, coherence, fluency and overall quality of summaries.

This evaluation focused only on quantitative methods and abstractive summarization models like BART and T5. These models allowed generating clear results regarding the validity of the new metrics. This method was important for discovering the pros and cons of the measures in assessing abstract summaries and helped in enhancing the evaluation techniques in this domain.

3.6.3 Data Participant Selection

The new metrics for automatic summarization of text were investigated through a quantitative research design. This method evaluated the metrics technically through measures focused on the statistical performance ensuring an objective and comprehensive evaluation.

In this study, the population for the quantitative design was the complete text summaries generated by several abstractive summarization models across various domains like a news article, scientific paper, etc. The benchmark datasets (e.g., CNN/Daily Mail) were used for

summarization to create the sample. To generate high-quality abstractive summaries, T5 and BART were used on these datasets for abstractive summarization.

The evaluation of the summaries generated by the various models was done using both the traditional metrics, such as BLEU, ROUGE, METEOR, BERTScore and new metrics developed. This way statistical assessment was done on how well the new metrics performed and compared to established standards in terms of meaning, coherence, fluency and overall quality of the summary, among others.

By using quantitative methods and abstractive summarization models BART and T5, this evaluation was able to yield crystal clear evidence about the utility of the new metrics. This method was important for realizing the merits and demerits of the metrics in evaluating quality of abstractive summaries, which helped improve the evaluation of abstractive summaries.

3.6.4 Instrumentation

This paper collected critical and essential data and assessed the performance of the new metrics for automatic text summarization using a quantitative approach. We gathered data from summaries produced by selected models and evaluate these summaries by using traditional and novel metrics.

- **Primary Data Collection:** The primary method used two types of metrics including traditional metrics and new ones to evaluate summaries made by abstractive models. To achieve this, we utilized the CNN/Daily Mail dataset which is a well-known benchmark for summarization. For summary generation, we chose BART and T5 models, which are known for their high performing models used in abstractive summarization. The assessment was considered for classic measurements like BLEU, ROUGE, METEOR, and BERTScore as well as the new metric. The metrics provided quantitative data for the summaries allowing direct comparison and statistical analysis

of the summaries. The evaluation process was automated using python libraries to apply these metrics on all summaries consistently. The collected quantitative data assisted in understanding how well the newly proposed metrics reflect the meaning, coherence, consistency and overall quality.

- **Secondary Data Collection:** In order to carry out the text summarization project, secondary data was collected by the help of research studies, articles, etc. We reviewed the prior work done on traditional metrics like BLEU, ROUGE, METEOR and BERTScore along with their shortcoming and advantages. Investigating the application of these metrics in past researches helped us get some insight that helped us improve them in evaluating the quality of abstractive summaries generated by models such as BART and T5.
- **Data Processing:** Python scripts were used to apply both traditional and new metrics to the generated summaries ensuring consistency and accuracy in the evaluation process. The results were stored in a structured database facilitating detailed statistical analysis. Comparisons between the performance of BLEU, ROUGE, METEOR, BERTScore and the new metric were made using statistical methods to assess how effectively each metric evaluates summary quality. Correlation analysis was conducted to see how closely the new metrics align with traditional metrics in capturing aspects like fluency, coherence and relevance.

This study evaluated the new metrics thoroughly and rigorously by solely using quantitative methods and compared several existing and new metrics. The aim was to give an objective, measurable picture of how these summarization metrics performed with respect to the quality of the generated abstractive summaries. Thus, we got a clearer understanding of their merits and demerits.

3.6.5 Data Collection Procedures

The data collection method in this study followed a structured and systematic approach with focus on the quantitative evaluation of summaries generated by abstractive models using traditional and new metrics. The process was divided into several below key stages:

- **Dataset Selection and Preprocessing:** The extensive CNN/Daily Mail dataset was our primary text source for generating summaries; popularly used benchmark for text summarization. This dataset was made up of news articles and human-written summaries perfectly used to test and evaluate auto summarization models. The data was fed to the models after processing like tokenization, normalization, filtering, etc., to ensure input format consistency. We used Python libraries like NLTK, SpaCy to preprocess and normalize text before producing summary.
- **Summary Generation Using BART and T5 Models:** We used both BART and T5 pre-trained models to create abstractive summary. Both models created summaries for each article in the preprocessed CNN/Daily Mail dataset. This was automated with the help of python library implementations (available via Hugging Face Transformers) (Wolf et al., 2020). The summaries generated were stored in a database (excel) in a properly structured form with the correct article summary generated matching to the original article for evaluation.
- **Application of Traditional and New Metrics:** Once the summaries were generated, they were evaluated using a set of traditional metrics, including BLEU, ROUGE, METEOR and BERTScore alongside the newly developed metrics. These summaries were compared against the reference human-written summary in the dataset using these metrics. Python libraries like NLTK, Rouge-score, BERT-SCORE and NLG-EVAL were used to compute the evaluation scores for each metric. The results were recorded in a database for statistical analysis.

- **Statistical Analysis of Metric Performance:** Once the metric scores were collected, the statistical analysis was conducted to assess the performance of the new metrics against the traditional metrics. The mean, variances and correlation T-test and Anova of different metrics were carried out using statistical tools. In statistics, we used significance testing to determine if the differences in performance are statistically meaningful or if they have happened by random. This study assessed how effective new metrics captured key aspects of summary quality, including coherence, meaning and consistency.

By implementing these procedures for data collection, the study objectively examined the new metrics for automatic text summarization in a systematic manner. We got detailed and measurable insights into the performance of the new metrics against the benchmark metrics.

3.6.6 Data Management

Data management in this study was handled in a systemized manner. We developed data management procedures that allowed effective storage and retrieval of input (summaries) and output (metric scores), while ensuring analysis and comparison and benchmarking is smooth.

- **Data Storage:** All the data including that of the original sentences from the CNN/Daily Mail dataset, generated summary by BART and T5 and the reference summaries were stored in excel. This excel listed below fields for:
    - **Source Text:** This field is the original article from the CNN/Daily Mail dataset.
    - **Reference Summary:** This field is the human-written summary paired with the source text.
    - **Generated Summary**: This field is the summaries generated by BART and T5 models.
    - **Metric Scores**: Scores generated by each evaluation metric (BLEU, ROUGE, METEOR, BERTScore and new metrics).

Data was stored in structured CSV/Excel files, making it scalable and easy to fetch. Additional copies of the database were routinely kept on external hard drives and google cloud to prevent data loss.

- **Data Organization and Structuring:** Data was presented to allow easy analysis across different metrics, models and articles. The goal was to have efficient analysis. Each record/row in the database included all the fields. This enabled horizontal (across metrics) and vertical (across summaries) comparisons. We also recorded the metadata like the model that got used (BART or T5) and the preprocessing applied to the text.
- **Data Backup and Recovery:** To avoid loss of data, database and other data files were regularly backed up. In the event of a system failure or data corruption, all of this would be kept on outside hard drives so that it could be restored quickly.
- **Data Cleaning and Validation:** Before conducting a final analysis, data cleaning was done to remove incomplete and corrupted records. This pertained to verifying if each of the source texts has a reference summary, generated summary and evaluation score. In case of any incomplete or wrong data entry, the same would be flagged. Validation scripts were run to check for data integrity to make sure the metrics were applied correctly to the summaries.
- **Data Retention and Disposal:** After the study finished, the data would be available for a certain period for further academic research, business review for studying the same thing again. Once the retention period expires, it will be deleted or anonymized as per the standards.

By adhering to the above data management methods/practices, the research ensured that all research data are secure, systematically filed and available for easy access-analysis. This paper ensured maintaining the integrity of it and reproducibility of findings.

3.6.7 Data Analysis

We analyzed the summary generation from the BART and T5 abstractive models using the CNN/Daily Mail dataset. The analysis primarily targeted the novel metrics along with the traditional metrics like BLEU, ROUGE, METEOR and BERTScore. This section gave an overview of the process of analyzing the input data, which comprises the source texts and the summaries, and evaluates them using these metrics which are then followed up by considered statistical comparisons and expert analysis.

1. Input Data Analysis

The CNN/Daily Mail dataset source articles and their BART and T5 model-generated summaries were the input data for this work. The next part of the work examined the properties of the input data which influences and affects their summarization performance.

- **Source Text Analysis:** The CNN/Daily Mail dataset consisted of news articles and corresponding human-written reference summaries. Each source text was analyzed in terms of length, complexity and word length. Descriptive statistics were calculated for the input articles:
    - **Word count** and **sentence length** of the source texts.
    - **Lexical diversity**: An analysis of vocabulary richness in the source texts which may influence summarization difficulty.

By learning the characteristics of source texts differ, it was easier to explain varying performance in summarization, particularly where certain types of articles (longer, complex, etc.) lead to better or worse summaries.

- **Generated Summary Analysis:** Summaries produced with the models, BART and T5 were analyzed to assess their structural and linguistic features, including below:

- **Length of summaries**: The number of sentences or tokens in produced summaries.
- **Compression ratio**: Ratio between the length of the original text and the length of the summary.
- **Repetition and redundancy**: Repetitive phrases/content to indicate model weakness in summarizing effectively.

Input data analysis helped to set the context for statistical analysis. For example, if a type of article or summarized output got better or lowered metric scores.

2. Quantitative Data Analysis

Post the input data was analyzed, the evaluation of the generated summaries using BERTScore, ROUGE-1, METEOR and BLEU along with new proposed metric was started:

- **Metric Evaluation Across Models:** Evaluation of Metrics across Models: Each summary generated by the models BART And T5 were put under traditional metrics like BLEU, ROUGE, METEOR and BERTScore as well as new metrics. The models' performances were compared on the basis of these scores:
  - **BLEU and ROUGE**: These metrics measured n-gram overlap between the generated and reference summaries which highlighted how well the model preserved key information.
  - **METEOR**: This metric assessed alignment in the meaning between the generated and reference summaries by incorporating synonym matching and stemming.
  - **BERTScore**: This metric evaluated semantic similarity using transformer-based embeddings, providing a deeper understanding of how well the generated summary captured meaning.

- **Descriptive Statistics of Input Data:** Descriptive statistics of the input data (e.g., word count, sentence count of the source texts) were correlated with the summary evaluation scores to determine if specific types of articles lead to better or worse performance under different metrics. For instance:
    - Long or complex articles may produce higher or lower scores depending on the metric.
    - Summaries with higher compression ratios may show different behavior in terms of BLEU or ROUGE versus more semantically-oriented metrics like BERTScore.
- **Model-Specific Performance Analysis:** A comparison was used to find how BART and T5 effectively summarized different categories of source texts. Both model's summaries were analyzed with metrics to determine what kinds of articles led one models to outperform other. Factors such as:
    - Length of input articles if one model is more suited for longer articles.
    - Summary length may indicate how well each model condenses content.

3. Statistical Comparison of Metrics

In order to assess the performance of conventional and new metrics and their capability of capturing the quality of summaries produced from the inputs, statistical measures were used:

- **Input Data Correlation:** Correlation analysis was conducted on input data features in relation to the cosine similarity scores for each BART and T5 summary from the original article. This analysis helped us find out whether certain features of the source texts characteristics affect the similarity between generated summary and source article. For instance, if the length of the article is high, the cosine similarity score may not be high, and thus it may reflect some bias.

- **Significance Testing:** Statistical tests (e.g., paired **t-tests** or **ANOVA**) were performed to assess whether the new metrics provide significantly different or improved evaluations compared to traditional metrics. These tests analyzed whether there are meaningful performance differences when evaluating summaries generated from models.

4. **Visualizing Data**

- **Data Visualization:** The generated summaries between BART and T5 were visually represented. The input data which was the length distribution of the source articles and source summaries was visualized. Box plots and histogram showed the relationship between input data properties.

3.7 Reliability and Validity of Study

Verifying the reliability and validity of study was an essential part in the evaluation of new measures of automatic summarization with the CNN/Daily Mail dataset with BART and T5 Large language models.

3.7.1 Reliability: Reliability refers to the consistency and stability of the results.

- **Consistency in Metric Application**: The study incorporated the CNN/Daily Mail dataset and applied the metrics (BLEU, ROUGE, METEOR, BERTScore and new metrics) uniformly to the summarization models (BART, T5). Python scripts evaluated metrics making sure that they are applied consistently and human error is reduced.
- **Reproducibility**: Summaries were generated using fixed model parameters and random seeds to guarantee consistent output. Test-retest reliability confirmed that the metrics produced stable results when re-applied under the same conditions.

3.7.2 Validity: Validity ensured that the study accurately measured the effectiveness of the new metrics.

- **Content Validity**: The metrics evaluated multiple dimensions of summarization quality such as accuracy and consistency. Both traditional and new metric were applied to ensure comprehensive evaluation.
- **Construct Validity**: Statistical analysis, such as correlation tests, t-test and Anova between traditional metrics (BLEU, ROUGE) and new metric, were established to check whether the new metric effectively measured the intended aspects of summary quality.

While the quantitative approach to evaluating new metrics for LLM-generated summaries using BART and T5 models offered a systematic method, it also came with certain limitations.

3.8 Research Design Limitation

Quantitative Research Limitations:

- **Dataset and Model Selection Bias**: The paper relied on the CNN/Daily Mail dataset and the used of abstractive models such as BART and T5. The performance of new metric on specific datasets or models may introduce unknown bias. This suggested that to achieve the optimal results, we may need to tune other hyper-parameters before trying for extractive summarization or dataset with longer contextual data parameters.
- **Focus on Quantifiable Data**: While the quantitative approach generated objective, reproducible data, it may miss the human judgment and aspects of the summaries like context and meaning which is generally difficult to measure. A more important element of summary quality may not be measured adequately by quantitative metrics.

3.9 Conclusion

In this chapter, study presented a quantitative methodology that was designed in order to assess new metrics for automatic text summarization. These summaries will be generated with large language model (BART and T5). Using a quantitative method, the present study aimed and objectively determined the scores of the traditional metrics (BLEU, ROUGE, METEOR and BERTScore) together with the new metrics developed in this paper. The research design included an extensive systematic plan for the collection and analysis of data. To conduct robust testing, the research leveraged a well-known dataset called CNN/Daily Mail. The quantitative method enabled a consistent evaluation where traditional and newer metrics were applied to summarized outputs. Statistical analysis was conducted to compare the performances. This methodology ensures the results were reliable, reproducible, and offered a detailed comparison of the metrics. The data management practices were clearly outlined to ensure how the data will be secured and how it will be handled throughout the experimentation. Also, the chapter gave key considerations for maintaining reliability and validity, emphasizing the importance of statistical rigor in ensuring the consistency and accuracy of the findings. In conclusion, this chapter sets the stage for the more detailed quantitative study of new summarization metrics. The strategy presented here is methodical which highlights their relevance and efficiency in enhancing the evaluation of automatic text summarization activity while also highlighting their behavior in comparison with the classic evaluation metrics.

Chapter 4 EXPERIMENTS AND RESULTS

4.1 Introduction

In the previous chapter, we presented research methodology based on which this project was undertaken. This chapter talk about the implementation of those methodology on the dataset to check performance of new metrics for automatic text summarization. For long time, metrics such as BLEU, ROUGE, METEOR and BERTScore have been commonly used to assess the quality of a summary. The advancement of AI technologies has caused these metrics to fail in some aspects of summary quality that require deeper consideration, such as meaning and coherence. To overcome the limitations discussed above, we applied statistical analysis to evaluate the performance of the new metrics objectively.

We used statistical means to evaluate how well the new metrics enabled the assessment of summary quality in terms of fluency, coherence and meaning. This method guaranteed that

assessment was based on quantifiable, repeatable data that provided an objective measure of the performance of the traditional as well as new metrics.

The implementation process was completed by applying the traditional metrics (BLEU, ROUGE, METEOR, BERTScore) as well as new metrics on the summaries generated by BART and T5. The CNN/Daily Mail benchmark dataset was used for evaluation in this study. By evaluating the outcome through comparatively different metrics, we collected details which show the strengths and weaknesses of the various evaluations.

The study carried out correlation analysis as well as descriptive statistics to examine the performance of these metrics. This enabled comparison of traditional measures against new ones that were clear and objective. The findings showed how well the new metrics captured nuances of meaning and coherence which the traditional metrics miss out on.

Because we rely mostly on statistics to draw conclusions so it helped us in focusing on the data which is objective and presentable. This saved the assessment of the novel metric from random results. The analysis provided a solid basis for evaluating the new metrics against traditional ones, as well as insight into their improving power of automatic text summary evaluation.

The next sections look at the steps taken in the process such as data collection, statistical analysis and how the result is presented. The chapter aims to illustrate how a purely quantitative approach through statistics made a mechanistic evaluation of the new metrics on BART and T5 summaries possible. In further sections, we encompassed all the metrics so as to contribute a lot more effective evaluation mechanism for automatic summarization.

4.2 Dataset Description

The CNN/Daily Mail dataset consisted of more than 300,000 news articles written by journalists at CNN/Daily Mail. These articles covered many topics in English Language. Initially, the dataset was developed for the aim of machine reading comprehension and

question-answering in an abstract way. The present version of the dataset can be utilized for two types of summarizations namely, extractive summarization which picks out important parts of the text and abstract summarization which rewrites the text and conveys the essence in a novel way.

4.2.1 Data Instances

The items presented in the dataset consist of three components, namely:

- ID: A unique identifier of the article.
- Article: Each article can be easily identified and accessed with this field. The news article provides full text that is the context for the summarization or question answering tasks.
- Highlights: A highlight is a short summary of the article often consisting of one or more sentences that served as the target output in the summarization task.

Like this, as an example:

- ID: 0054d6d30dbcad772e20b22771153a2a9cbeaf62
- Article: A recent article revealed that an American woman died on a cruise ship. The particular ship also reported several sick passengers. The cruise ship's name is MS Veendam. The woman is said to have suffered from serious health complications which resulted in her death.
- Highlights: The highlights reveal that the elderly woman had diabetes and hypertension and that previously, there was a case of 86 passengers falling ill.

The average token counts for the articles and highlights in the dataset were as follows:

- Article: 781 tokens
- Highlights: 56 tokens

4.2.2 Data Fields

The fields of the dataset were as follows:

- ID is a string in the hexadecimal formatted SHA1 hash of the URL.
- An article is a string containing the full text.
- The highlights are the summary of what the article writer has written in the article.

4.2.3 Data Splits

The CNN/Daily Mail dataset was split into three main sections mentioned below:

- Train involved 287,113 instances
- Validation involved 13,368 instances
- Testing involved 11,490 instances

These splits were very important to develop and evaluate summarization models. These were tested on unseen data to evaluate model performance.

4.3 Dataset Creation

4.3.1 Curation Rationale

The CNN/Daily Mail dataset has evolved through several versions, each aimed at addressing different research needs.

Version 1.0.0 supports supervised neural techniques for machine reading and question answering. It has a little over 3,13,000 unique articles and close to 1 million questions.

Versions 2.0.0 and 3.0.0 shifted attention from question answering to summarization. The dataset was updated to allow summarization tasks in version 3.0.0 (now non anonymized). This is the main change in this version which is contrasting to previous versions which anonymized named entities.

4.3.2 Source Data

CNN/Daily Mail dataset articles (April 2007 – April 2015) were taken from their respective archives (of www.cnn.com and www.dailymail.co.uk) as archived by Wayback Machine. Version 1.0.0 couldn't process documents beyond 2000 tokens; hence those articles were

removed. All the data can also be obtained from other sources such as Github. It has the new code for the non-anonymized data. Also, latest version has the further changes. Tokenization and normalization of the data were handled using specific scripts. The previous version 1.0.0 (Hermann et al., 2015) script was used for earlier versions. See's (See et al., 2017) tokenization script of Version 3.0.0 was adopted. This script lower cases text and adds missing periods.

4.3.3 Source Language Producers

The dataset has names especially from Version 3.0 which gives a more realistic attribution of the language used in the articles. But this dataset doesn't contain any information about the original authors of the articles.

4.3.4 Considerations for Using the Data

The CNN/Daily Mail dataset is designed for building models that can summarize long articles into short summaries. This is one of the most used datasets for text summarization experiments. Though selection process helped to convey information/data quickly from lengthy texts; one should be aware that anything produced by models trained on this data will tend towards the language of articles (which may be biased) and will not always reflect the refinements of the content.

4.3.5 Discussion of Biases

According to Bordia and Bowman (2019), the CNN/Daily Mail dataset exhibits some gender bias, but as compared to other dataset it is less. Also, the dataset's perspectives are (mostly) from the US and the UK which could influence their take on world events. Kryściński et al. (2019) highlight that news stories usually offer essential information in the first few lines, so may be this affects the distribution of information in summaries. Chen et al. (2016) mentioned

that many of the samples were so ambiguous and messy that they were challenging even for human annotators.

4.3.6 Other Known Limitations

Machine-generated summaries can vary in truthfulness compared to the original articles. While extractive summarization models aim to present accurate extracts, discrepancies may arise between the summaries and the actual content of the articles. This detailed examination of the dataset's structure and creation provided a detailed understanding of its components and foundation behind its curation and the considerations for its use in summarization research. Also, it was worth to note that the articles were written by and US and UK people with specific event related to people in the US and the UK, hence most of data and language may be related to that.

4.4 Data Understanding

The initial dataset has a training, validation and test dataset with three columns id, article and highlights which are all non-null objects.

- **The training set** has 287,113 entries in it, with each article and highlight having a unique identifier.
- **The validation set** consists of 13,368 entries with unique articles and highlights.
- **The Test set** has 11,490 entries containing almost all unique articles and highlights.

Original Dataset Characteristics

- **Training Set:**
    - Contains total unique articles worth 284,005 (some are repeated).
    - Total unique highlights: 282,197 (83 duplicate highlights).
- **Validation Set:**
    - All 13,368 articles are unique.

- 13,300 are unique highlights, while 16 are duplicate highlights.

- **Test Set:**
  - Most entries are unique in the test set. Only 2 articles are repeated and 3 highlights are duplicated.

Missing Data

None of the datasets had missing values so they are complete and do not need imputation.

Test Set Usage in This Study

In this study we only analyzed the test set. Since we are using an already trained Large Language Model (LLM), we do not need to train or validate it any further. So, the training and validation datasets are not needed. Also, an evaluation requires little hyperparameter tuning.
We used the complete test set of 11,490 rows to comprehensively evaluate the performance of the model.

4.5 Data Analysis

The focus of this study was on the test dataset only for model evaluation. However, analyses on the training and validation datasets were also performed. This method also guaranteed that all datasets have similar means and standard deviations. A thorough examination of the training, validation datasets revealed that the test dataset has the characteristics and patterns of the entire dataset. Hence, in experiments only the test dataset was used. Since as analysis, it was observed that test dataset captured all the major features and aspects of the data. Having this understanding helped validate our results, thus making our assessments more reliable.

4.5.1 Significance of word count distribution in Article:

The word count distribution in articles was important for several purposes; NLP tasks like summarization, classification or any other model dealing with textual data fall under this.

Thus, due to overall growth in articles, it was important to analyze the count. Being aware of the average article length helped with model engineering, hyperparameter choices and data pipeline optimizations. Balancing datasets help the model see input of different sizes to learn efficiently and effectively. This analysis showed how article length differs across the datasets (train, validation, test) which in turn helped model to learn accordingly.

Summary of "Article Word Count" Distribution:

This analysis showed the distribution of word counts in articles in the train, val and test dataset. All the distributions presented in each dataset were visualized using a histogram with 50 bins as well as a smooth Kernel Density Estimate (KDE) curve to facilitate identification of patterns.

1. **Train Dataset Distribution:**

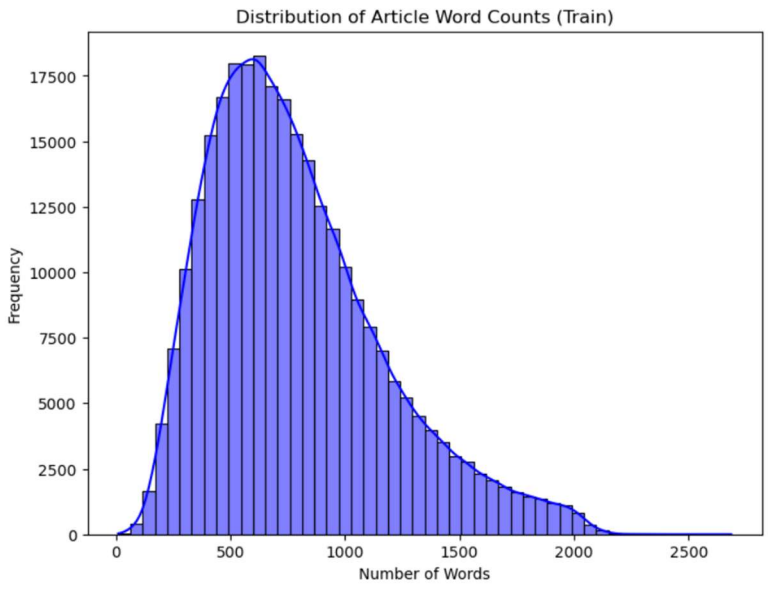

Figure 4.1 Distribution of Article Word Count (Train)

- Observation: The graph was right-skewed, it indicated that most articles had low words whereas only few articles had high words.

- Peak: Most of the articles were of size 400-600 words. This size had about 17500 articles.
- Spread: The majority of articles had word counts between 100 and 1,500 words, though some went beyond 2,000 words.
- Long Tail: A small number of articles had more than 2,000 words, but they are very rare.

2. Validation Dataset Distribution:

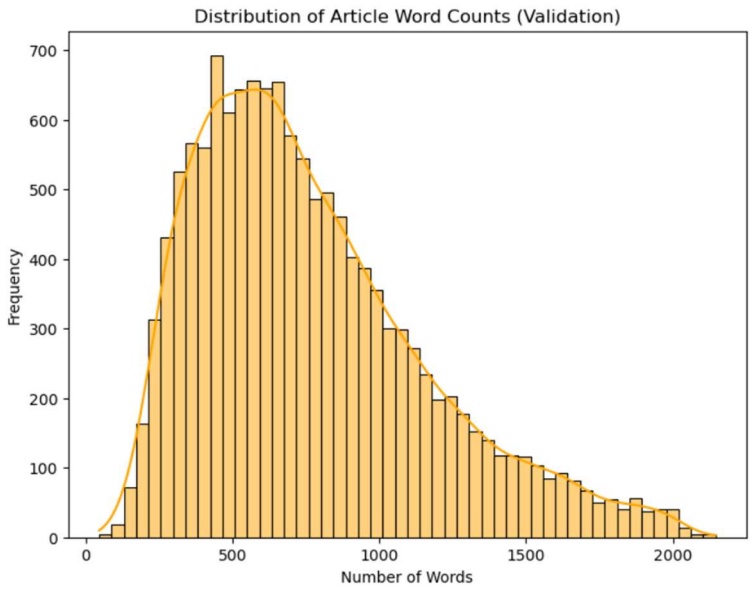

Figure 4.2 Distribution of Article Word Count (Validation)

- Observation: Like the train dataset, the validation set was also right-skewed.

- Peak: Most articles were again in the 400–600-word range, with around 700 articles at the peak.
- Spread: Most articles had between 100 and 1,500 words, similar to the train set.
- KDE Slope: The number of longer articles reduced after 1,500 words.

3. Test Dataset Distribution:

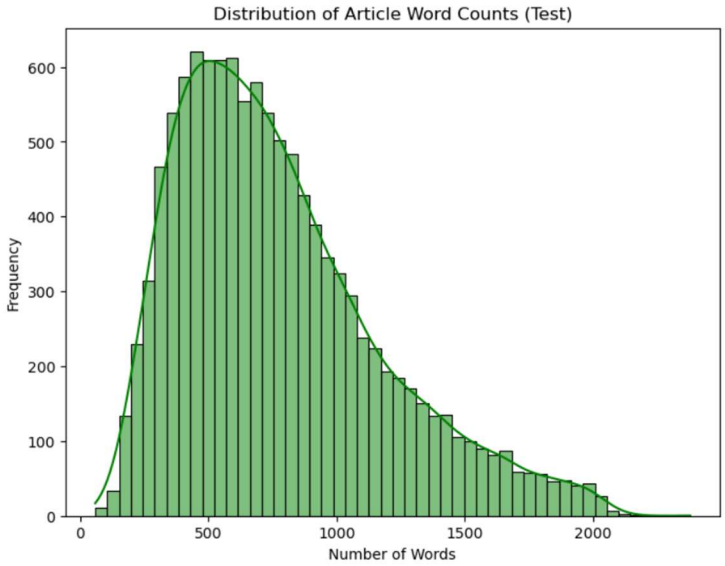

Figure 4.3 Distribution of Article Word Count (Test)

- Observation: The distribution of test set was also right skewed as we see in other two sets.
- Peak: Most articles were between 400 600 words with a peak frequency of around 600.
- Spread: Most articles were falling in the 100-1,500-word range.
- Long Tail: There were very few articles with more than 2,000 words, similar to the train and validation sets.

Overall Insights:

- All three datasets showed a right-skewed distribution, where most articles were around 400-600 words.
- The long tail was composed of the much rarer items like the one at the right end of the curve.
- The train, validation and test sets had very similar word count distributions making it easier to train and test models consistently.

4.5.2 Importance of Analysis word count distribution in Summary:

It was important to see the word count distribution of the summaries that will help us to design and optimize the machine learning model. We used the distribution of texts to design a text summarization model. This helped in understanding the nature of the data that LLM model will get for training by looking the summary lengths. We also checked if these distributions are similar across the train, validation and test sets. This helped in optimizing the model for shorter inputs if dataset has summaries with short text. If there were lengthy abstracts, then the model should also learn more complex data. This aided in adjusting the model to fit the requirements of the dataset.

Summary of "Summary Word Count" Distribution:

In this kind of analysis, visualization of the distribution of the summary word count of the test, train and validation datasets was done. The train set had a wider variety, including longer summaries and similar trait were observed in test and validation datasets, that helped model to understand the nature of the data.

1. **Train Summary Word Count Distribution:**

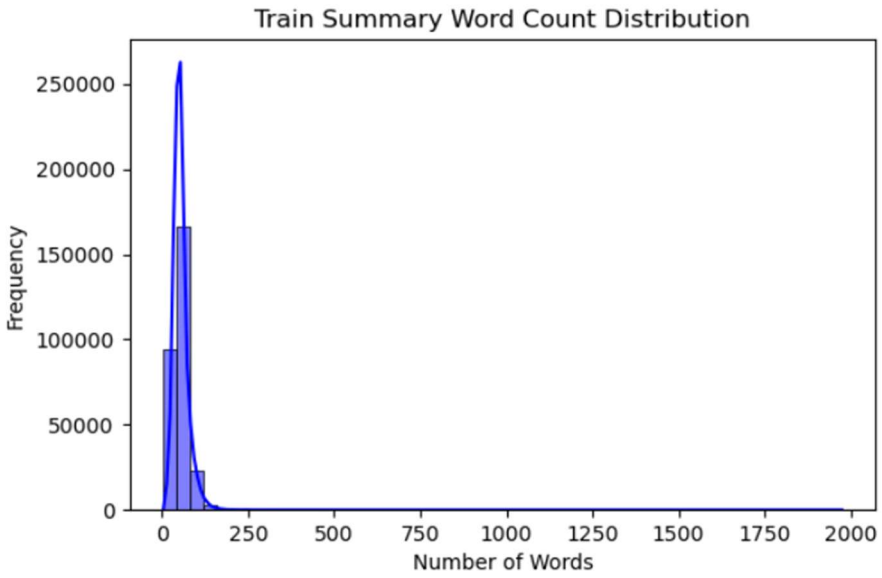

Figure 4.4 Distribution of Summary Word Count (Train)

- Shape: The data was skewed to the right, meaning that many of the summaries were short and only a few are long.
- Peak: Most of the summaries in the training data were on the shorter side. Most of the summaries contained between 0 to 50 words.
- Range: Some summaries went upto 1,000 words, though that were very rare. The number of summaries diminishes significantly post the 200 words.
- KDE Line: The KDE line was smooth indicating that many short summaries existed but very few long summaries.

2. **Validation Summary Word Count Distribution:**

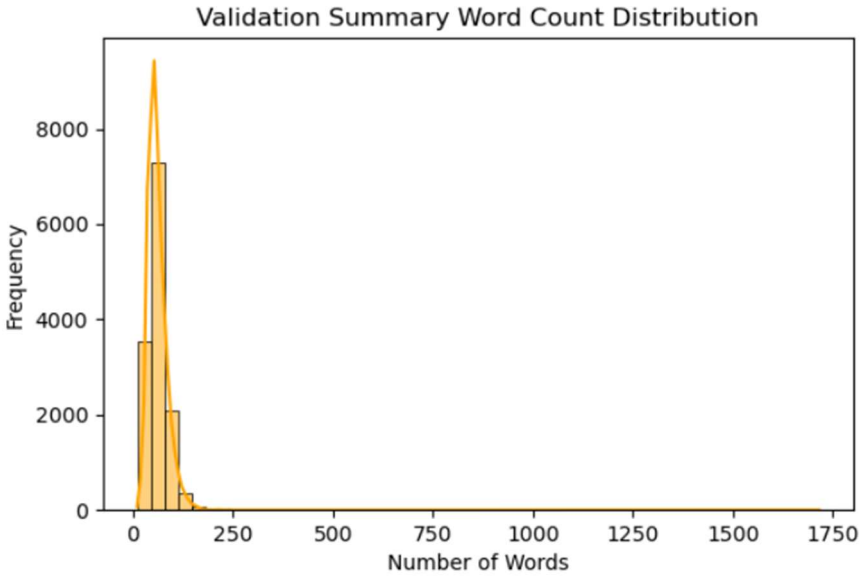

Figure 4.5 Distribution of Summary Word Count (Validation)

- Shape: This also had a right-skewed shape, similar to the training data, but the peak was sharper.
- Peak: Most summaries in the validation data had 20 to 50 words. This meant that validation summaries were also quite short.
- Range: A few summaries had more than 500 words, but those were very rare. Most summaries were less than 100 words.
- KDE Line: The line again showed that most summaries were short, with a sharp decrease after 50 words.

3. Test Summary Word Count Distribution:

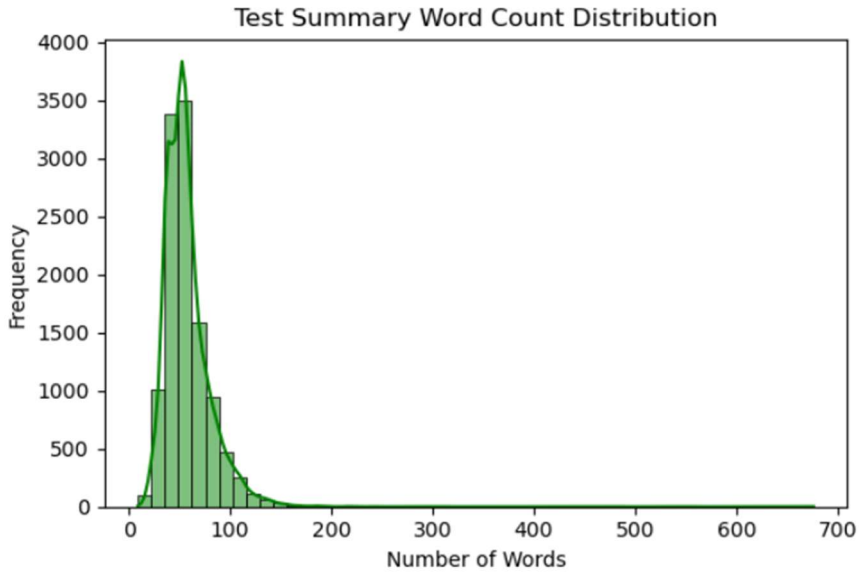

Figure 4.6 Distribution of Summary Word Count (Test)

- Shape: The test data was also right-skewed, but the skewness was less extreme than the other two datasets.
- Peak: The peak was around 20 to 50 words, meaning most test summaries were also short.
- Range: Hardly any test summaries were longer than 300 words, and the majority of the data was less than 100 words.
- KDE Line: The line shows that most test summaries are short, with fewer long ones.

Overall Insights:

- Right Skewness: All three datasets showed right-skewed in plot that means most summaries were short, with very few long ones.
- Train vs Validation vs Test:
    - The train set had more variety in summary lengths. Some summaries were much longer (over 1,000 words).
    - The validation and test sets had shorter summaries on average. Most were below 200 words.

- Distributions: The train set had many more data points, so it showed a wider range of word counts. The validation and test sets had fewer summaries and were more focused on shorter ones.

4.5.3 Importance of Analyzing sentence count distribution in Article:

The distribution of sentence count in articles was important for the purpose of developing and fine-tuning the machine learning model for text summarization task. By observing the distribution of sentences in train, validation and test sets in visualization plots, we learnt that the model can be tuned to generate the summary basis these distributions. If articles are short, the model can be adapted to small inputs. But if there are lengthy then the same complex longer sentences data can also be adapted. We studied these distributions to ensure that the datasets we selected for training and testing were balanced. This ensured optimization across the distribution of article lengths.

Summary of Article Sentence Count Distribution:

This study sought to analyze the distribution of Article Sentence counts in training, validation and test set. To avoid a biased model, the training and validation data used should equally represent the lengths of documents as are available in test data. We can optimize the performance of the model such that it performs well on both long and short articles by knowing how many sentences each article usually has in the datasets. The following showed how many sentences were in each of the three datasets.

1. **Train Article Sentence Count Distribution:**

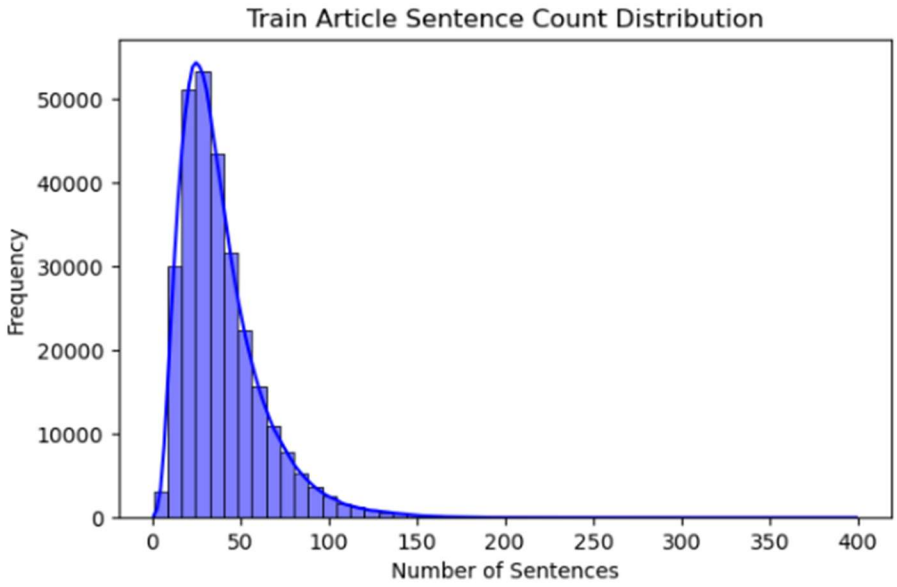

Figure 4.7 Distribution of Article Sentence Count (Train)

- Sentence count distribution of train was right skewed meaning most of these articles had less sentence and smaller number had many sentences.
- Most articles were about 25 to 30 sentences long in total.
- Articles contained anywhere from 0 to over 400 sentences, though there were not many articles that were longer than 100 sentences.
- Conclusion: Training set offered a wide range of lengths of articles and they were long enough to allow the model to learn effectively.

2. Validation Article Sentence Count Distribution:

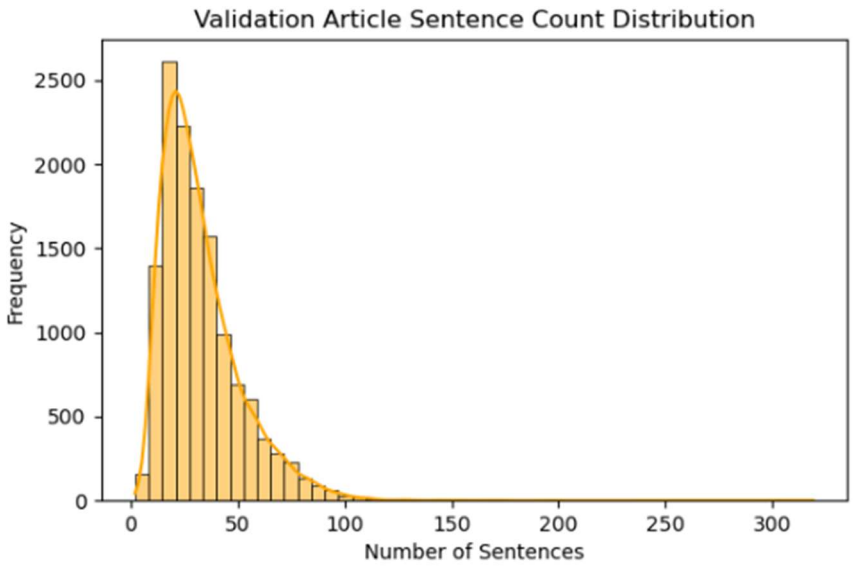

Figure 4.8 Distribution of Article Sentence Count (Validation)

- The validation set distribution was right-skewed, yet it was more heavily concentrated on shorter articles.
- Most articles contain 15 to 20 sentences.
- The articles had between 0 and 300 sentences, with the vast majority having under 100 sentences.
- Conclusion: The validation set was designed to pick mostly smaller articles, allowing hit and trial faster in training time and allowing better generalization of the model.

3. Test Article Sentence Count Distribution:

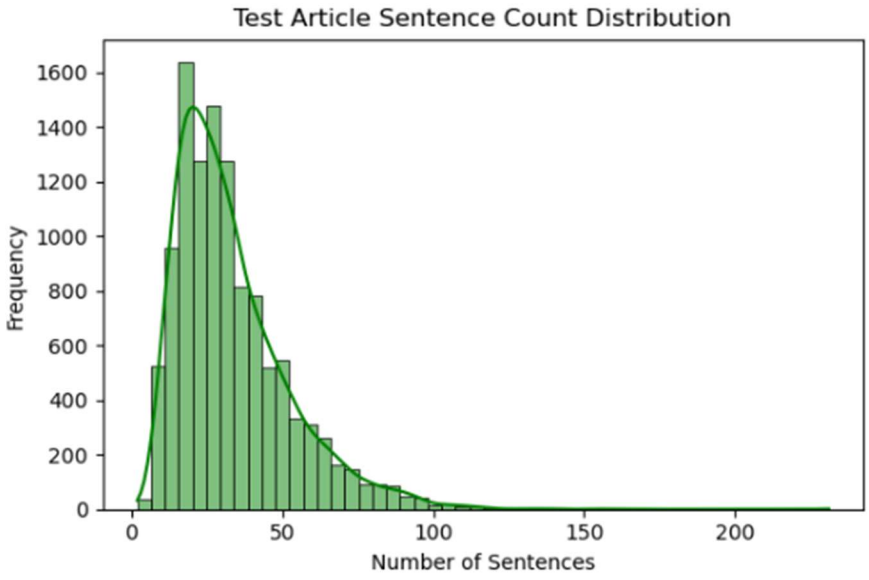

Figure 4.9 Distribution of Article Sentence Count (Test)

- The test distribution was also slightly right-skewed like our training distribution.
- Most of the articles consist of 20 to 25 sentences.
- The test set of sentences was similar to the training set when it comes to the quantity of sentences found. That was, the counts range from about 200 all the way to 0.
- Conclusion: Therefore, the distribution of the test set was a good approximation of the training set.

Overall Insights:

1. **Training Set Diversity**:

   The training set features a wide range of sentence counts where the sentences were picked from a variety of technology-related topics. Most articles had anywhere between 25 and 30 sentences. Out of all, the one used 160 times was quite popular and had hundreds of thousands of views. Nevertheless, a smaller section was assigned to very long articles (up to 400 sentences).

2. **Validation Set Focus:**

The focus of the validation set was specialized. It was on shorter articles. Most of these articles had sentences between 15-20. This allowed efficient evaluation during training phase. Validation articles were not that long.

3. **Test Set Similarity to Training**:

   Test set was similar to train set in the number of sentences – distribution of test set sentences was quite similar to train set, with the most popular being 20-25 sentences. Testing was done on data that was a lot like training data and validation data hence safe to assume the test data exhibited similar trait and could be used directly in our experiments. Also, the similarity ensured that the model was tested on a data set which the model was already trained to.

4.5.4 Importance of Analyzing Summary Sentence Count Distribution:

It was important to analyze the distribution of the summary sentence count for tasks such as summarization, abstraction and generation among different articles. It was essential for the model to know how many sentences a summary would generally have so that it was adjusted to the dataset. This review allowed for preprocessing and model design alternatives and guarantees the model to permits for differing length summaries within each dataset.

Summary of "Summary Sentence Count" Distribution:

As seen below, the article summaries in the train, validation and test datasets were examined for their distribution of the number of sentences, as seen here. The charts showed the number of sentences in each summary of the different datasets and how they were scattered.

1. **Train Dataset Summary Sentence Count Distribution:**

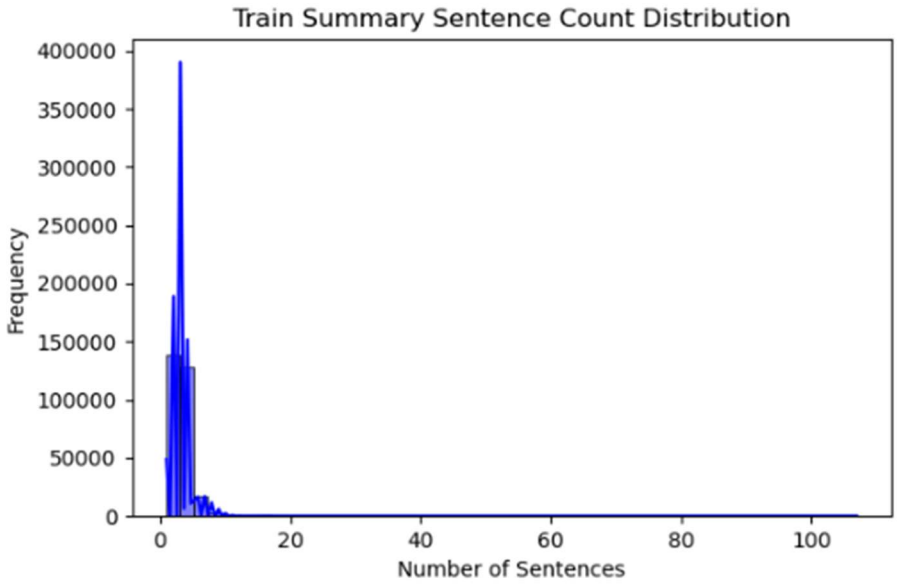

Figure 4.10 Distribution of Summary Sentence Count (Train)

- Observation: The distribution was highly right-skewed, with most summaries containing very few sentences.
- Peak: Most of the summaries in our dataset had been 1-5 sentences long, with peak frequency occurring at around 400000 summaries having 1 sentence.
- Long Tail: It was rare, but a few summaries did have 50+ sentences, which created a visible tail. Most people had a small negative sentence.

2. Validation Dataset Summary Sentence Count Distribution:

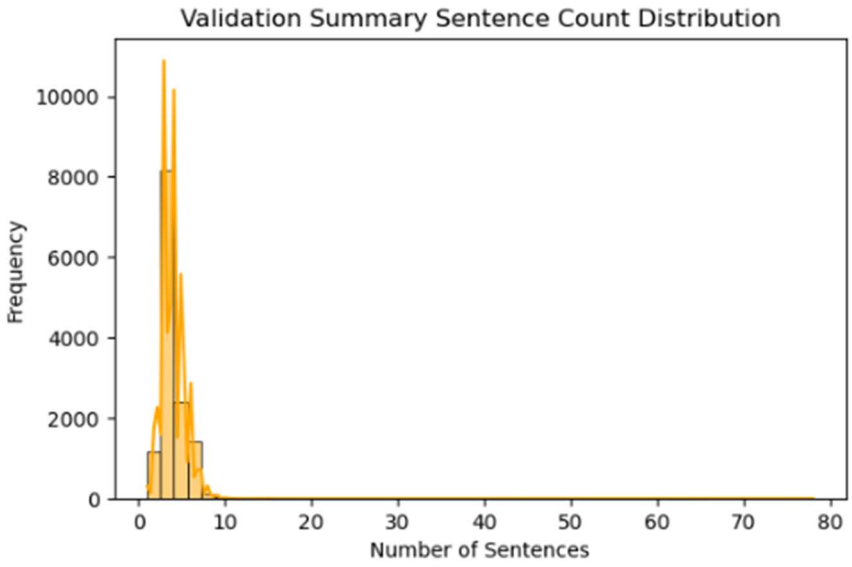

Figure 4.11 Distribution of Summary Sentence Count (Validation)

- Observation: Similar to the training set, the validation set was also right-skewed in nature as a majority of the summaries also have less number of sentences.
- Peak: Most summaries were 1 to 5 sentences long. The peak frequency of summaries was around 10,000 of which 2 sentence summaries peak.
- Long Tail: Some summaries exceeded 50 sentences long. They were rare examples. Most summary lengths were under 10 sentences or 1-10.

3. Test Dataset Summary Sentence Count Distribution:

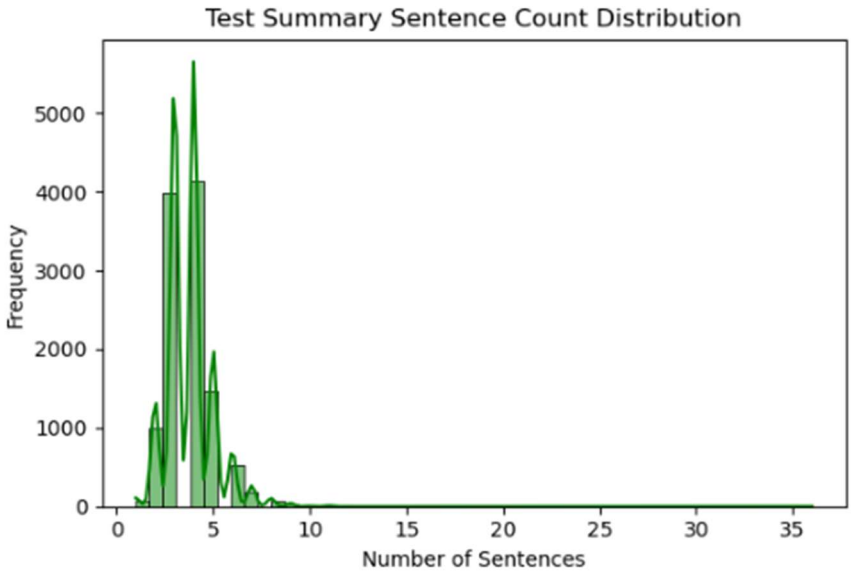

Figure 4.12 Distribution of Summary Sentence Count (Test)

- Observation: The test set also had a right skewed distribution and the summaries were all short in length.
- Peak: Most summaries had a peak of about 2 to 5 sentences. Peak frequency occurred for about 5,000 summaries with 4 sentences each.
- Long Tail: Although there were some summaries with more than 30 sentences, they were extremely rare and most summaries falling within the 1-10 sentence range.

Overall Insights:

- Most articles were short, according to all 3 datasets. All the three distributions of sentence count of summaries were right skewed.
- In all datasets, the maximum number of summaries with 1-5 sentences were seen and frequency rapidly dropped for longer summaries.
- All datasets exhibit a long tail, which indicated that most summaries were executed in the short length-range. But certain articles had a much longer summary length. However, these were uncommon.

4.6 Summary Generation Using an LLM-Based Algorithm

Summarization was an important task in natural language processing. It decreased an article into small form. For example, if we took a news article, we can condense the full text without losing the important information within. This study evaluated BART and T5. They are powerful models that can generate summaries. Here, we specifically used news as content. Hugging Face and Google had trained huge neural networks on tons of data that can summarize any text easily because they can understand the context of the text pretty well.

4.6.1 BART (Bidirectional and Auto-Regressive Transformers)

The BART model was engineered to perform exceptionally well on language tasks and summarization. It used encoders and decoders, bidirectional and autoregressive context to get better performance. Let's analyze that in more detail.

1. **Architectural overview**

Encoder:

- Bidirectional Reading: The encoder reads the input text from left to right as well as from right to left. In other words, it can look at the meaning of a word with the help of the meaning of every other word in the text.
- Transformers: The encoder consists of multiple transformer layers which are built with self-attention mechanisms. These layers help the model focus on different parts of the text dynamically.

Decoder:

- The decoder creates a summary word by word due to its autoregressive generation. The next predicted word is based on the words it has already generated and the encoded information from the article. This is done using masked self-attention to ensure that the model doesn't peek at future words while generating the current word.

- Causal Language Modeling of decoder is similar and resembles to human writing one word at a time that ensures coherence and fluidity in the final summary.

2. Training Part of Denoising Autoencoder

- Pre-training Phase: BART is trained as a denoising autoencoder in pre-training phase. At this point in time when the input text is corrupt intentionally (like masking or shuffling words of the sentences) the model learns to restore the text. This training teaches BART how language works so it can make a summary that is intelligible and relevant to the context.
- Fine-tuning Phase: BART is adjusted to perform a specific task for which it is fine-tuned for summarization. Fine-tuning helps BART learn how to summarize effectively.

3. Summary Making Process Phase

- Tokenization: The input article is first tokenized or logically separated into manageable pieces to make the model understand.
- Encoding: The tokenized article is further fed into the encoder where it produces a set of hidden states which capture the article's context.
- Decoding: The decoder takes these hidden states and starts generating the summary word by word until it reaches the specified length or an end-of-sequence token.

4.6.2 T5 (Text-to-Text Transfer Transformer)

T5 is one of the methods to summarization and other natural generation tasks. The main idea is to make the job a text-to-text task so that it’s easy to train. Let's take a closer look at its architecture.

1. Architecture Overview

Unified Framework:

- The T5 framework is based on the premise that any text can be transformed into another text, regardless of its nature. In other words, regardless of whether it is

translating a sentence, answering a question or summarizing an article, the same thing is happening.

Encoder-Decoder Structure:

- Encoder: T5 also features an encoder-decoder architecture with the encoder resembling that of BART. The encoder reads the entire input sequence and processes everything to create a representation with rich context. The transformer layers in the encoder use a stack. Thus, it is able to use self-attention on different text parts.
- Decoder: The output text of the decoder is obtained by using transformer layers; however, it is obtained one by one. T5's Decoder makes use of the output generated by the encoder and creates the summaries word-by-word making sure that the words fit well.

2. Pre-training with a Denoising Objective

- Text-to-Text Training: T5 receives training on a variety of tasks and used this training to model text to text in a generalized manner. Hence; we see T5 as Text-to-Text transformer.
- Denoising Objective: Similar to BART; T5 is trained with a denoising objective: portions of the input text are masked, and T5 learns to predict them. It shows AI how language is put together and helps AI with writing to make it better.

3. Process for Summary Generation followed in T5

- Encoding Part: This part helps tokenized text goes into the encoder that produces a series of hidden states that encapsulate the context.
- Decoding Part: The decoder combines those hidden states to produce a summary one word at a time or until the end of the summary.

4.7 Hyperparameters in Summary Generation

Hyperparameters are settings that control how models like BART and T5 summarize text. Tuning the hyperparameters influenced the quality and the speed of the generated summaries. The main hyper-parameters for each model are mentioned below.

1. **Hyperparameters for BART**

- Max Length
  - Definition: The maximum tokens allowed in the summary being generated is called its max_tokens limit.
  - Impact: If the value is too high, the notes can become long and lose focus, while if the value is too low, information may get omitted.
- Min Length
  - Definition: Minimum tokens needed in the summary is referred to as min length.
  - Impact: To make sure that the summaries do not lose important content by not becoming too short.
- Length Penalty
  - Definition: A longer summary will incur a penalty when being scored.
  - Impact: A longer summary will incur a penalty when being scored. When the length penalty exceeds one, it will discourage longer outputs. When it is less than one, it encourages longer outputs balancing Conciseness & Informativeness.
- Number of Beams
  - Definition: The number of beams used in the beam search algorithm for generating summaries.
  - Impact: More beams allow the model to explore multiple possible summaries before selecting the best one. While this generally improves quality, it also increases computational demands.

- Early Stopping
    - Definition: A mechanism to stop summary generation once the model produces a satisfactory output.
    - Impact: This can help save processing time by preventing unnecessary additional tokens from being generated.

2. Hyperparameters for T5

- Max Length
    - Definition: Similar to BART, this hyperparameter denotes the upper limit on the number of tokens in the generated summary.
    - Impact: It is essential to tune with care so that the summary is not too long and nor too short.
- Min Length
    - Definition: The minimum length that the summary must meet.
    - Impact: This helps ensure that the summary is not too brief, conveying the message of the article.
- Length Penalty
    - Definition: It means the quantity of beams utilized in T5's decoding process.
    - Impact: helps to balance between the shortness and the lengthy content of the output.
- Number of Beams
    - Definition: The number of beams used in T5's decoding process.
    - Impact: Increasing the number of beams can lead to better output quality, but more beams will also require more computation. Hence it is most important to balance.
- Early Stopping

- Definition: This flag indicates when the generation process should stop based on reaching an end-of-sequence token or sufficient output length.
- Impact: It helps improve efficiency by cutting off generation once a satisfactory summary is created, saving processing time.

Conclusion on Hyperparameters

The hyperparameters for both BART and T5 were important for controlling the quality and performance of the output.

With power of fine tuning, model can generate both informative and concise summaries that accurately reflecting the article's main points. In summary tasks, the configurations we used greatly affected the quality of the summary. Thus, it was important to experiment with the configurations to fetch the best performance.

4.8 Details of the Generated Summary

This study investigated the performance of BART and T5 state-of-the-art models on the CNN/Daily Mail dataset containing original news articles, human written summaries and machine generated summaries. This was a quick analysis of the BART and T5 summaries produced accurately. Important measures used in the analyses were word count distribution, cosine similarity between human summaries and model generated summaries. The analysis of word count distribution was helpful to compare the model's performance on shortening of articles with that of human produced summary. The usage of cosine similarity helped us understand how similar or dissimilar are the original summaries from the summaries generated by the machine. It gave us to assess how faithfully BART and T5 captured the core information from the articles. By comparing the results generated by both models, we tried to understand which model is better.

Dataset Overview and Structure

The dataset consisted of the below important columns:

- **Original Article**: The text of the original article.
- **Original Summary**: Human-created summary about this section.
- **Summary by BART**: The summary generated by the BART model.
- **Summary by T5**: The summary generated by the T5 model.
- All columns were complete with no missing data.

1. Word Count Analysis

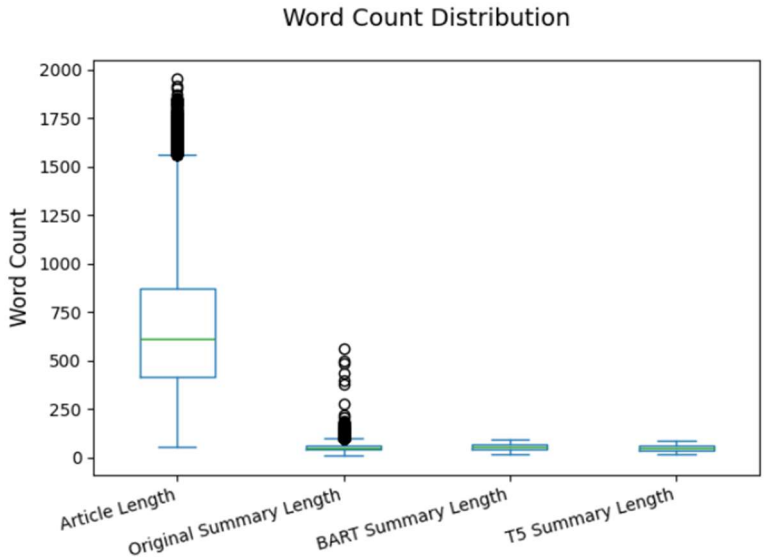

Figure 4.13 Distribution of Word Count in Generated Summary

The articles and summaries vary greatly in length. Thus, we presented below a summary of the lengths of the articles.

- **Article Length**: The original articles had a mean length of about 679 words but varied widely from a minimum of 70 words to a maximum of 1821 words, as shown by the large spread in the box plot.

- **Original Summary Length**: The previous summaries were shorter with an average word count of 55 words. According to the box plot, there were some longer summaries but the max length can be 500. However, most summaries were short.
- **BART Summary Length**: Summaries produced by the BART model were pretty short, consisting of only 54 words on average. Very few outliers were present in the distribution.
- **T5 Summary Length**: Summaries produced by the T5 model were slightly shorter than the BART summaries, with an average of 49 words, and the distribution showed fewer outliers than BART.

Key Observations:

- **Article Length Variance**: The original articles vary greatly in length as shown by the large interquartile range (IQR) and outlier presence, meaning that some articles were substantially longer than their peers.
- **Summary Length Consistency**: In contrast, both the human-written summaries and the model-generated summaries (BART and T5) show much tighter distributions.
- **BART vs. T5**: The human-written summaries and model-generated summaries (BART and T5) had much tighter distributions than length consistency exhibits.

  T5 generated summaries that were shorter on average than BART summaries. Thus, both models generated concise summaries but T5 was doing even better.

This data was visualized in the box plot, showing:

- **Variance in article lengths**: Some articles were much longer than others, as indicated by the large range.

- **Tighter summary distributions**: Both the BART and T5 summaries had more consistent lengths, with the T5 model producing slightly shorter summaries on average compared to BART and original summaries.

2. Cosine Similarity Between Summaries:

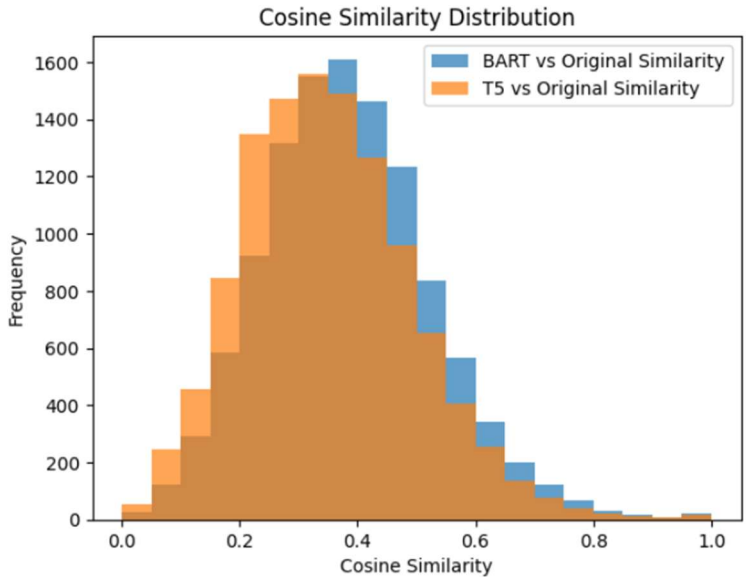

Figure 4.14 Distribution of Cosine Similarity for Generated Summary v/s Original Summary

To find the closeness of summaries generated by the machine (BART and T5) to that of the original human-written summaries, cosine similarity was computed. Similarity scores can be 0 or 1 where 0 means no similarity and 1 means perfect similarity.

Key Results:

- BART Summaries: The average cosine similarity between BART-generated summaries and the original human-written summaries is 0.385, indicating that BART summaries capture a substantial portion of the same content as the original summaries.
- T5 Summaries: T5 summaries were slightly less similar to the original summaries, with an average cosine similarity of 0.350, suggesting a marginally lower overlap with the original content compared to BART.

The cosine similarity distribution histogram revealed:

- Most of the BART and T5 summaries had similarity with scores between 0.2-0.5 range. Which means they showed partial overlap with the original summaries.
- BART exhibited a higher frequency of scores towards the upper end of this range, particularly between 0.4 and 0.5, suggesting it had a slight advantage in capturing the original summary content more closely than T5.

4.9 Experiments carried with generated summary with Traditional Evaluation Metrics

The efficiency of automatic summarization systems were measured with the help of different quantitative metrics. This experiment used a multitude of evaluation metrics to assess summary quality, incorporating BLEU, METEOR, ROUGE (ROUGE-1, ROUGE-2, ROUGE-L variants) and BERTScore-F1. With the different strengths and weaknesses of each metric, we used a weighted average to combine these scores for easier evaluation of the performance of summarization.

Summarization Metrics

1. BLEU (Bilingual Evaluation Understudy)

This metric evaluated the n-gram overlaps count captured in the generated summaries and reference texts, primarily focusing on precision. measuring how many words sequences in the summary matched the reference exactly. The scores from our experiments were:

- **BART:** 0.1441
- **T5:** 0.1071

The efficiency of automatic summarization systems can be measured with the help of different quantitative metrics. With the different strengths and weaknesses of each metric, we used a weighted average to combine these scores for easier evaluation of the performance of summarization. The higher BLEU score for BART suggested that it performed better in generating summaries with a closer lexical match to the reference texts. However, BLEU's

reliance on exact word matches made it less sensitive to variations in wording or deeper semantic meaning, which may explain its moderate scores overall.

2. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

This evaluation metric was an improvement over BLEU since it took precision and recall into account. It considered synonym and word stems unlike BLEU which does not consider. The METEOR scores were:

- **BART: 0.3766**
- **T5: 0.3312**

The BART model had higher METEOR scores than the T5 model which showed that it was more lexical rich as well as content-rich. Compared to T5, BART gave more extensive and richer semantic summaries.

3. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE worked on recall and measured the n-grams overlap (i.e. how much of the reference content was included in the generated summary):

- **BART had**:
    - **ROUGE-1**: 0.4374
    - **ROUGE-2**: 0.2099
    - **ROUGE-L**: 0.3079
- **T5 had:**
    - **ROUGE-1:** 0.4044
    - **ROUGE-2:** 0.1833
    - **ROUGE-L:** 0.2833

BART consistently surpassed T5 in all ROUGE metrics, suggesting that it retained a higher amount of important information from the reference texts. According to ROUGE-1

(unigram overlap) and ROUGE-L (longest common subsequence) scored for BART, BART better retains important terms and overall content structure.

5. BERTScore-F1

Unlike traditional n-gram-based metrics, BERTScore evaluated the semantic similarity between generated and reference summaries using contextual embeddings. It captured more nuanced relationships and deep meaning beyond word-for-word matching. The BERTScore-F1 results were:

- **BART:** 0.2883
- **T5:** 0.2241

The higher BERTScore for BART indicated that its summaries not only aligned better lexically but also captured deeper semantic similarities with the reference summaries. This showed that BART generated summaries that were both informative and semantically coherent.

Combining Metrics with Weights

We used a weighted score mechanism that gave more importance to metrics measuring semantics and lexical diversity, as opposed to simple word overlap measures, to get a complete picture of the model's abilities. The weights for each metric were defined as follows:

- **ROUGE-1**: 0.4
- **METEOR**: 0.3
- **BERTScore-F1**: 0.3

By combining these metrics, we emphasized content retention (ROUGE-1), lexical variation (METEOR), and semantic fidelity (BERTScore), which together provided a more well-rounded evaluation of summarization quality.

Rationale for Weight Assignments

- **ROUGE-1 (0.4)**: This measure received the highest weightage of 0.4 owing to the fact that the retention of content is quite significant in summarization. High ROUGE-1 scores reveal that the summary comprises important content..
- **METEOR (0.3)**: The METEOR evaluation metric (0.3) takes into account the precision and the recall along with the synonyms and the word stems. This measure was highly weighted because it better captured linguistic variation, indicating a more natural and flexible use of language.
- **BERTScore-F1 (0.3)**: BERTScore was weighted equally with METEOR because it captured the semantic similarity between generated summaries and reference texts. As the evaluation of meaning and coherence was essential for effective summaries, this weight reflected its importance in modern summarization evaluation.

Combined Score Calculation

The ultimate bundled score for each model was determined through the application of respective weights as per the guidelines discussed above on the particular metric scores of the respective models and summation of the results.

- ROUGE-1: 0.4374
- METEOR: 0.3766
- BERTScore: 0.2883

Combined Score (BART)= 0.3744

The same goes for T5, which had:

- ROUGE-1: 0.4044
- METEOR: 0.3312
- BERTScore: 0.2241

Combined Score (T5) = 0.3283

So, BART's combined score = 0.3744, and T5's = 0.3283.

Results and Discussion

The experimental results using traditional metrics (BLEU, METEOR, ROUGE) and also BERTScore all showed that BART outperforms T5:

- **BART** outperformed on every standard metric, whether BLEU, METEOR or ROUGE, showing it preserves content and lexical similarity better.
- The **BERTScore** results showed how well BART manages to maintain semantic fidelity, which showed that its summaries are not only correct but also more coherent and faithful to the references.
- The combined scores clearly indicated that the gap between the models was large. BART scored 0.3744 as compared to T5's 0.32783. Thus, it showed a more balanced performance.

This evaluation suggested that BART generated more informative, coherent, and contextually accurate summaries, making it a stronger model for summarization tasks however the evaluation metrics in totality performs lower than 40%.

Conclusion

In this assessment, we applied n-gram reliant metrics (BLEU, METEOR, ROUGE) along with semantic reliant evaluation (BERTScore) in order to measure summarization models performance. No matter the metric used, BART outperformed T5 on every one of them. Even with BART's higher performance, all scores remained below 40%, which means although

BART generated better summaries in terms of lexical accuracy and semantic meaning, the quality of the overall evaluation metrics can still be improved.

The multi-metric weighted approach combined multiple metrics to take care of accuracy on the surface level and beyond that still, the low scores suggested a mismatch between the evaluation metrics and their limitations in evaluation. This showed that we need to develop the evaluation metrics to solve more complex tasks.

4.10 Experiments carried to evaluate generated summary with wordnet based similarity scores

In addition to traditional evaluation metrics like BLEU, METEOR and ROUGE, we also leveraged WordNet-based similarity metrics in our experiments to address some limitations of these conventional approaches. Unlike traditional metrics which mostly focus on lexical matching and n-gram overlapping, our metrics captured concept similarity which included semantic relationship and concept hierarchies. To overcome this limitation, we investigated WordNet-based metrics, including the Average Path Similarity, Average Normalized LCH Similarity, Average WUP Similarity, and Average Lin. This helped us measure not just word similarity but also the measurement of words and concept's semantic and hierarchy relations.

1. **Jaccard Similarity**

- **Definition -** Jaccard Similarity is a statistical metric that tells us how similar two sets are. The size of the intersection divided by the size of the union of the sets is called the Jaccard Similarity.
- **Methodology -** The sets A and B used in the study referred to the words/phrases in the generated summary and in the reference summary, respectively. The Jaccard Similarity gives the extent of matching in the vocabulary and hence an easy measure of lexical similarity.

- **Word Usage:** While Jaccard Similarity helps to know how much words are used in the summary generation; it might not fit semantically. This is since Jaccard Similarity only checks the matching words or phrases.

2. **Traditional Cosine Similarity**

- **Definition:** Cosine Similarity measures the cosine of the angle between two non-zero vectors in a multi-dimensional space. It determines how similar the two vectors are irrespective of their magnitude.
- **Methodology:** In summarization, vectors are usually built using term frequency-inverse document frequency (TF-IDF) or word embeddings. The cosine similarity is a metric used to measure how similar two vectors are. The measure finds the dot product of the vectors and divides it by the magnitude (length) of each vector. The result of the cosine similarity measure is any real value between -1 and 1.
- **Word Usage:** The metric is effective at capturing the total relation between two summaries due to the direction of vector as opposed to just their counts. The higher the score, the more similar the contexts of the summaries.

3. **Average Path Similarity**

- **Definition:** Average Path Similarity measures the semantic distance between concepts based on their paths in a semantic network or ontology.
- **Methodology:** In this metric, we compute the average path length between words or phrases in a graph representation of the ontology (like WordNet). The shorter the path, the higher the similarity.
- **Usage:** This measurement is really useful for evaluating the relationships between ideas that are mentioned in text summaries. However, it can depend on the quality of the underlying ontology.

4. **Average Normalized LCH Similarity**

- **Definition:** The Average Normalized Lowest Common Hypernym (LCH) Similarity refers to the measure of similarity based on the lowest common hypernym of a taxonomy.
- **Methodology:** The closeness of the two concepts is based on the distance between the concepts in a taxonomy like WordNet normalized by the depth of the taxonomy.
- **Usage:** LCH Similarity highlights the relationships between concepts in a hierarchy. It allows us to see how well the summary made captures the themes and relations present in the reference summary.

5. Average WUP Similarity

- **Definition:** Wu and Palmer Similarity (WUP) measures the similarity between two concepts based on their depth in a taxonomy and the depth of their least common ancestor.
- **Methodology:** WUP Similarity which is a measure for the similarity of the two concepts based on the depth of the concepts themselves. Moreover, also based on the closest common ancestor of those concepts in the hierarchical structure, yielding a score that reflects how closely related the concepts are.
- **Word Usage:** This metric captures semantic relatedness more effectively than Jaccard Similarity which is less context-aware. Higher WUP scores mean the summary still has meaningful connections with the reference.

6. Average Lin Similarity

- **Definition:** Lin similarity shows how similar two concepts are. To determine how similar the concepts are, Lin similarity uses information content. Furthermore, Lin similarity uses the least common ancestor of the two concepts.
- **Methodology:** Below formula is utilized to calculate similarity.

$$\text{Lin Similarity (A,B)} = \frac{2 * IC(LC\ A(A,B))}{IC(A) + IC(B)}$$

where IC is the information content derived from the corpus used to build the taxonomy.

Word Usage: The Lin's Similarity considers the information content of the terms. It provides an in-depth analysis of the semantic similarity of a generated summary with respect to the information content of the reference summary.

4.10.1 Experiments with Data

In addition to traditional summarization metrics, evaluating the semantic similarity between generated summaries and reference texts was crucial for understanding the quality of summarization models. The research tried Jaccard Similarity, Traditional Cosine Similarity, Average Path Similarity, Average Normalized LCH Similarity, Average WUP Similarity and Average Lin Similarity measures along with others. Each metric gave various perspectives on how closely the generated summaries match the original pieces of content. We were checking if weighted average of these metrics could give a robust framework for a comprehensive evaluation.

Similarity Metrics

1) Jaccard Similarity

The Jaccard Similarity was the measurement of similarity between two sets which can be computed by taking their intersection and union. It was a particularly helpful measure to check how many unique words or phrases the summary shares with the reference.

Jaccard Similarity scores we achieved:

- **BART**: 0.25
- **T5**: 0.22

Both models had similar Jaccard scores indicating a fair amount of overlap.

2) Traditional Cosine Similarity

Classic Cosine Similarity measured the cosine of the angle between two non-zero vectors which took place in an inner product space. It was used to measure orientation and not magnitude. This similarity measure helped to understand the semantic relations between two text vectors. These vectors were created with help of simple count vectorizer that check presence or absence of word in matrix. The scores were:

- **BART**: 0.51
- **T5**: 0.48

This suggested that BART's summaries were more semantically similar to the references than T5's summaries.

3) Average Path Similarity

The semantic structure of the sentences was taken into consideration while calculating Average Path Similarity. It checks the average differences and seperation between ideas in a semantic network. The scores were:

- BART: 0.15
- T5: 0.16

Both models exhibited almost similar results that reflects their comparable ability to maintain relevant conceptual relationships.

4) Average Normalized LCH Similarity

The average normalized LCH similarity worked using the Lowest Common Hypernym (LCH). LCH similarity was defined based on the path length between concepts in a taxonomy. The results were:

- BART: 0.24
- T5: 0.25

BART and T5 results revealed that T5 creates better extraction and cover larger area.

5) Average WUP Similarity

The Average WUP Similarity was based on Wu and Palmer where the similarity of two concepts is calculated by taking into consideration their depth in a taxonomy and the depth of that ancestor on wordnet. The results were:

- **BART**: 0.32
- **T5**: 0.33

Both models produced high scores, with T5 again showing a marginal advantage.

6) Average Lin Similarity

Average Lin Similarity used the information content of concepts and their least common ancestor in a taxonomy. The scores achieved were:

- **BART:** 0.12
- **T5:** 0.13

A higher score on T5 indicated a better representation of information content of the summaries.

Combining Similarity Metrics with Weights

We used weighted combinations to analyze the evaluation results of the semantic similarity metrics. With this strategy, we stressed some of the similarities more than others, depending on what was being summarized. Also, the weights for each metric were defined as:

- **Jaccard Similarity**: 0.05
- **Traditional Cosine Similarity**: 0.25

- **Average Path Similarity**: 0.10
- **Average Normalized LCH Similarity**: 0.05
- **Average WUP Similarity**: 0.30
- **Average Lin Similarity**: 0.25

Rationale for Weight Assignments

The weight assigned to the metrics showed their importance in semantic similarity assessment as below

- **Jaccard Similarity (0.05)**: A lower weight was assigned due to focus on lexical overlap rather than deeper semantic relationships.
- **Traditional Cosine Similarity (0.25)**: A moderate weight was given to this metric as it captured the similarity of generated and reference summary at great extent.
- **Average Path Similarity (0.10)**: This measure was important but is given a lower weight than others based on hierarchical semantic ordering.
- **Average Normalized LCH Similarity (0.05)**: Due to its niche use and low coverage, it had a relatively low weight.
- **Average WUP Similarity (0.30)**: This evaluation metric took into account the conceptual hierarchy and similarity based on hierarchy hence gave more weight since it was more useful for summarization.
- **Average Lin Similarity (0.25)**: Lin Similarity was given the higher weight as it was looking at capturing vital measure of information content.

Combined Similarity Calculation

The combined similarity score for the summary was estimated by finding the weightage of each WordNet similarity score for each of summary and then multiplying them. The goal of

this method was designing a comprehensive evaluation method to assess the semantic quality of the summaries produced by the systems. The WordNet-based similarities capture some semantic meaning, the score mostly remained around 40%.

The combined WordNet similarity scores for BART and T5 were as follows:

- **BART Combined WordNet Similarity**: 0.29
- **T5 Combined WordNet Similarity**: 0.29

The scores indicated that BART performed better than T5 on WordNet-based semantic similarity. Still, since both got below 40%, it showed that only using WordNet Similarity will not be sufficient for evaluation.

Results and Discussion

The weighted combination approach equally assessed the WordNet similarities for both the summarization systems. Both models did not capture the semantic relationship of words as their combined scores for any individual WordNet similarity measure was less than 40% indicating that even their similarity scores did not correspond well.

Conclusion

This section talked about the WordNet similarity metrics that are used to measure the performance of the summarization, and gave details about the chosen weights for each. The weighted combination approach forces the similarity value given by the wordnet based similarities. As a result, a broad assessment of the semantic quality of the combinations can be made. However, the scores of combinations remain near 40%, indicating the limitations of the WordNet similarities as evaluation metrics. Although they captured deeper relationships between words; more complex evaluation techniques and improvement of the metrics should be developed for automatic summarizers for better handling of semantics.

4.11 Methodological Triangulation for Summary Evaluation

We propose a new summarization evaluation metric USES (Unified Summary Evaluation Score) that used methodological triangulation to combine traditional evaluation metrics and more advanced semantic metrics to overcome the traditional metrics and wordnet based metrics. This approach combined the strengths of both metric types to create a more robust evaluation of generated summaries, addressing the limitations of using either method in isolation.

Traditional evaluation metrics, ROUGE-1 and METEOR were two most commonly used metrics to check the quality of summaries generated by machine learning models. They compared lexical similarity of extracted summaries with reference summaries. The metrics measured n-gram overlap, precision, recall and lexical similarity. For example, ROUGE-1 was a recall-based score that measured the overlap of n-grams from the reference summary to the candidate summary. Moreover, METEOR score included both precision and recall. Also, it considered stemming and synonym matching. While these metrics are effective at detecting a similarity in surface structure, they were lacking the capacity to capture similarity in deeper meaning or overall meaning of the summaries. Currently summaries that seem different in wording but mean the same are likely to get less scores in traditional metrics. Thus, evaluation processes looked incomplete.

To overcome the limitations mentioned above, we included metrics such as BERTScore, WordNet-based metrics (WUP Similarity), Cosine Similarity metric and Cosine Similarity using MiniLM (paraphrase embeddings) in our evaluation. These metrics go beyond surface-level word matching focusing on the semantic relationships between words and concepts. The generated summary was referred to as the reference which was very useful in knowing whether the generated summary gave the same message as the reference summary though the exact words might be different. BERTScore used contextualized embeddings of transformer-

based models to measure similarity between the semantic meanings of generated and reference summaries. Similarly, assessment of conceptual relationship between the terms was measured by WordNet based WUP Similarity on taxonomy-based parameters. Also, WUP Similarity based on WordNet on taxonomy based parameters was measured in order to assess the conceptual relationship between the terms. On the contrary, the embedding-based Cosine Similarity was used to determine the sentence-level semantic similarity of MiniLM by paraphrase embeddings.

We presented a composite summarization evaluation metric USES for evaluation of extractive as well as abstractive summarization. By combining traditional metrics (ROUGE-1, METEOR) and advanced semantic metrics (BERTScore, WUP Similarity, traditional Cosine Similarity and MiniLM-based Cosine Similarity), USES, captures both lexical and semantic meaning ensuring consistency and robustness. By considering both the literal and metaphorical use of words, this new measure offered a broader assessment. This composite metric served as a more useful measure of model performance in our experiments. For example, BART performed well on traditional metrics, indicating strong lexical overlap, while also excelling in semantic metrics, confirming that the generated summaries are semantically aligned with the reference texts. Also, despite the less lexical overlap of T5, higher performance on semantic metrics suggested T5 captures the gist of the reference summaries despite having a different choice of words.

This new composite metric enabled validating the quality of summarization models from different dimensions. Traditional metrics ensured that the generated summaries have the same words and structure as the reference summaries, while advanced semantic metrics ensured that the generated summaries had the similar meaning and were capable of producing a semantic similarity. When we used our new metric to evaluate summaries; it made sure that summaries are lexically sound and at the same time semantically accurate.

Using methodological triangulation, this new summarization evaluation metric not only addressed the limitations of existing metrics, which often overlook semantic fidelity, but also added the quantitative precision of traditional word overlap metrics to the semantic metrics. Both the basic text and the meaning of the produced summary would be evaluated using a word vector space method. This triangulated approach can therefore be considered an innovative solution for summarization tasks as it captures the entire range of summary meaning and form.

4.12 Summary

In this chapter, we analyzed the CNN/Daily Mail dataset in detail. We explained the working of the two state-of-the-art summarization models - BART and T5. Finally, we presented a new evaluation metric that used the conventional and WordNet metrics. The purpose of the current chapter was to determine the capabilities of these models through detailed experiments in summarizing the information and their performance using a combination of metrics and to introduce a new metric through methodological triangulation.

Dataset Characteristics

We started analyzing the CNN/Daily Mail dataset, which is a large-scale dataset for the summarization task. It contained the news article which can be used for abstractive summarization. With its size, diversity of content and well-structured format, it was suitable for evaluation of neural models in summarization. In this section, we have discussed the essential characteristics of the dataset which included the average length of the article and summary, the complexity of the languages and the type of information in the summary. Understanding the dataset's characteristics provided essential context for evaluating how well models like BART and T5 can condense, preserve information and how new evaluation metrics can be designed.

Understanding BART and T5 Models

The section explored the BART and T5 models, state of the art text generation large language models. BART, a type of Denoising Autoencoder called a Bidirectional and Auto-Regressive Transformer, helped to summarize the text and achieve abstractive summarization by generating fluent as well as coherent text. On the other hand, T5 (Text-to-Text Transfer Transformer) was a unified framework that views every NLP task as a text-to-text task and is capable of performing various text generation tasks like summarization. In this chapter, we also discussed the architecture and hyper-parameters of the models. This helped to understand how we can further use these models in our research paper to generate the summary.

Experiments with Traditional and WordNet-Based Metrics

We employed a few classical evaluation metrics like BLEU, METEOR and ROUGE to evaluate the performances of BART and T5. The metrics above usually measured how much a generated summary overlaps with a reference summary in terms of n-gram precision, recall and F1 Score. Although traditional metrics were useful for rough evaluations, they often do not account for semantic similarities between the source text and generated summary, especially in the case of abstractive summarization tasks. To overcome this issue, we proposed “WUP Similarity”, “Lin Similarity” and “Cosine Similarity” which are WordNet based metrics. The metrics look at the words and phrases in the generated summaries and reference summaries and see how close or far they are in meaning. We tried to develop a more useful metric by using both. It not only marked how much of the wording is being retained, but also how much of the meaning and essence was being retained from the original text in the summaries.

Weightage Assignment

In order to integrate these diverse metrics into a cohesive evaluation framework, we assigned specific weightages to each metric. Traditional metrics like BertScore, METEOR and ROUGE were given weightages based on their ability to measure lexical accuracy, while WordNet-based similarity metrics received higher weightage due to their focus on semantic content and conceptual alignment. This approach allowed us to place greater emphasis on meaning preservation which was often more important in abstractive summarization, while still considering lexical precision. We developed a composite evaluation score by assigning suitable weightages to reflect consistency and semantic quality of the summaries generated.

Methodological Triangulation

The final section of the chapter introduced methodological triangulation as a novel approach to summarization evaluation. We were able to cross-validate outcomes from various perspectives by integrating traditional metrics and wordnet-based similarity measures in one comprehensive metric. This triangulation method reduced the individual weaknesses of each type of metric, such as the over-reliance of traditional metrics on exact word matches and the neglect of structural precision by semantic metrics ensuring that both lexical accuracy and semantic coherence were accounted for. Thus, this new evaluation metric provides a more reliable and holistic assessment of model performance, capturing the full spectrum of summarization quality.

Chapter 5 DISCUSSION

5.1 Discussion of Results

The table below compared USES with the standard metrics ROUGE-1, METEOR and BERTScore for the BART and the T5 model. The evaluation was done through t-tests, ANOVA and Pearson correlation which represented two important things.

Table 5.1 Results

Metric	t-statistic	p-value	USES Std Dev	Metric Std Dev	Pearson Correlation	Correlation p-value	ANOVA Stat	ANOVA p-value	USES Better ?	More Consistent ?
BART_ROUGE-1	41.067	0	0.107	0.129	0.95	0	1686.519	0	Yes	Yes
BART_METEOR	74.09	0	0.107	0.146	0.857	0	5489.306	0	Yes	Yes
BART_BERTScore	128.919	0	0.107	0.142	0.862	0	16620.11	0	Yes	Yes
T5_ROUGE-1	36.684	0	0.112	0.132	0.945	0	1345.743	0	Yes	Yes
T5_METEOR	79.114	0	0.112	0.14	0.851	0	6259.054	0	Yes	Yes
T5_BERTScore	141.804	0	0.112	0.142	0.852	0	20108.5	0	Yes	Yes

- **Performance (Better?)**

The t-test compared the mean scores of USES with others to see whether USES gives better results. If the t-statistic is positive, it means USES is superior. However, when it is negative, that means the given metric is better than USES. P-values denotes the significance of differences.

- **Consistency (More Consistent?)**

  The ANOVA test evaluated the variance in the scores for each metric to understand how consistent USES metric was relative as compared to others. Lower standard deviations for USES suggested greater stability, and a significant ANOVA result (p-value < 0.05) indicated that this consistency was statistically meaningful.

Also, Pearson correlation values were given to test how much USES and each traditional metric corresponded. If something has a high correlation, it probably had the same trend as USES and may be better or more consistent. This comparative table on USES was greatly helpful in determining the effectiveness of USES against various other metrics.

The table above summarized USES measurements versus other measures in terms of effectiveness and reliability. The results showed that USES wins over ROUGE-1, METEOR and BERTScore for most comparisons (especially for both the BART and T5 models). The results of t- tests showed that USES achieved a much higher score than the other metrics. The ANOVA showed that USES was more consistent because it had less variance than ROUGE-1, METEOR and BERTScore. Also, the Pearson correlation values were high flipping from 0.851 to 0.950. USES was in good agreement with traditional measures but was more effective and stable. USES was a much more reliable and better metric to evaluate summaries made by large language models.

Discussions:

1. **T-Test Results: Why USES is Often Better**

The means of USES and other standard metrics (like ROUGE-1, METEOR and BERTScore) were compared with t-test. A high positive t-value indicated USES is better than the traditional metrics. This outcome was significant as per the shown p-value.

The BART and T5 scores on ROUGE-1, METEOR and BERTScore were again remarkably different from those of other models as the score differences were considerably high and were all significant as all the p-values were nearly zero.

- **BART:**
    - BART_ROUGE-1 (t-statistic = 41.067, p-value = 0.000) showed that USES is better than ROUGE-1.
    - BART_METEOR (t-statistic = 74.090, p-value = 0.000) showed USES is better than METEOR.
    - BART_BERTScore (t-statistic = 128.919, p-value = 0.000) showed USES has improved performance.
- **T5:**
    - T5_ROUGE-1 (t-statistic = 36.684, p-value = 0.000) showed significant improvement of USES over ROUGE-1.
    - T5_METEOR (t-statistic = 79.114, p-value = 0.000) highlighted superior performance of USES.
    - The T5_BERTScore result (t-statistic = 141.804; p-value = 0.000) confirmed USES was better than BERTScore.

2. ANOVA Results: Why USES is More Consistent

ANOVA tests showed that USES had lower standard deviations compared to other metrics, making it more stable. BART and T5 showed a large difference in variability of USES and

conventional metrics; also, USES's predictions were more stable. BART and T5 showed a large difference in variability of USES and conventional metrics; also, USES predictions were more stable than conventional.

- **BART:**
    - USES outperformed in consistency across all comparisons for other metrcis, with lower standard deviations (e.g., BART_ROUGE-1: USES std = 0.107 vs. ROUGE-1 std = 0.129).
    - BART_METEOR and BART_BERTScore also displayed similar lines trends, where USES demonstrated higher stability based on lower standard deviations and significant ANOVA results.
- **T5:**
    - T5_ROUGE-1 showed a significant improvement in consistency for USES (T5 std = 0.112 vs. ROUGE-1 std = 0.132).
    - USES was also more consistent across T5_METEOR and T5_BERTScore, as evidenced by their ANOVA results.

3. Correlation as an Additional Insight

Other measures with Pearson correlation coefficients ranging from 0.851 to 0.950 indicated USES generally performed better than the traditional measures but it was also picking up the same trends that the traditional measures do. Thus, USES can be said to be a better and stronger measure.

Conclusion

T-tests and ANOVA results suggested that USES consistently outperformed the traditional metrics (ROUGE-1, METEOR and BERTScore) in terms of performance and consistency. The USES metric was a better one for summarization evaluation of large language models due to its greater stability along with enhanced performance.

5.2 Discussion of Research Question One:

How can we evaluate summaries produced by large language models using different metrics or evaluation strategies not limited to BLEU and ROUGE?

The conventional metrics such as BLEU, ROUGE, BERTScore and METEOR that auto-evaluate summary, had their own limitations. The measures we're looking at focus on words and phrases overlapping on the surface. They thus fail to measure the semantic meaning of summaries and their coherence or informativeness. This was especially the case when a model makes paraphrase or employs other sentence structures. Here is how USES (Unified Summary Evaluation Score) fixed it:

- BLEU and ROUGE used the n-gram overlap for scoring. Though they calculated precision and recall, they did not account for semantic difference and word reordering still means the same. It led to lower scores for the summaries which were accurate but rephrased. USES solved this by integrating more sophisticated techniques for semantic evaluation.
- BERTScore used pre-trained BERT embeddings to evaluate how similar the words in a generated summary are to that of the reference summary. Yet, BERTScore mainly aptitudes for context-level similarity of words so it largely ignored overall communication and coverage. USES used BERTScore too, but they added other metrics measuring the coherence and meaning of sentences to improve on it.

- The METEOR metric had been designed to rectify the problems found in the earlier metric measures like BLEU and ROUGE. It achieved this by taking synonyms, stemming and paraphrasing into account while computing the score. However, this metric also failed at scoring a summary when it does not share much of its wording with the reference text. USES got over this shortcoming by combining the model with different evaluation layers to measure accuracy at different levels.

How USES Overcomes These Limitations

- **USES** employed traditional metrics as well as more sophisticated semantic and coherence evaluation methods to provide a better evaluation of summary quality. Below described are the core components of USES and how it improved evaluation:
- **Wu-Palmer Similarity (WUP):**
    - **How WUP Works:** By measuring the semantic distance between words, based on the hierarchy of the words in some lexical database like WordNet. When the generated summary and the reference summary contain different wording but similar meaning, this can help capture the relationship.
    - **Advantage:** With WUP in place, USES can properly evaluate summaries that make use of synonymy or paraphrase. It doesn't punish accurate summaries that use different words.
- **Cosine Similarity with Sentence Embeddings:**
    - **How It Works:** Cosine similarity compares vector representations of the whole sentence to determine how something will work. By this technique, USES

evaluated the degree of semantic similarity between sentences of two summaries.

- **Advantage:** When the USES measured similarity, it does not just use word overlap. The advantage of this is that; even if the summaries are expressed in a different way have the same meaning, it will not be a problem. It provided a more nuanced assessment of semantic accuracy.

- **Sentence Transformer Embeddings (e.g., Sentence-BERT):**
  - **How It Works:** These metric underlying uses Sentence-BERT and similar pre-trained models, USES created dense vector representations of entire sentences / paragraphs. We compared the embeddings to see how well the summary captures the meaning of the reference.
  - **Advantage:** USES had a benefit that picks-up coherence at sentence-level and paragraph-level; that is, a summary only requires not just a collection of accurate sentences, but also a text which made logical sense and was fluent.
- **BERTScore (Integration in USES):**
  - **How It Works: USES** included BERTScore as part of its evaluation toolkit, which focused on contextual meaning between words and their similarity based on pre-trained BERT embeddings.
  - **Advantage:** While BERTScore was good for checking meaning but USES was even better because it had extra measures that made it more stable and consistent. t-test and ANOVA showed that it works better.
- **ROUGE-1 and METEOR (Integrated in USES):**
  - **How It Works:** Besides advanced semantic evaluation strategies, **USES** also computed traditional metrics ROUGE-1 and METEOR for surface accuracy assessment.

- **Advantage:** USES incorporated a combination of traditional metrics and advanced semantic metrics so that n-gram scoring (ROUGE-1, METEOR) and deep semantic similarity were taken into consideration. This healthy approach had a balanced performance which leads towards better consistency of results as noted from the statistics.

With the use of some metrics like ROUGE-1 and METEOR along with more new metrics like Wu-Palmer similarity, sentence embeddings and cosine similarity; USES proved to be a reliable metric that can be used to evaluate summaries from large language models. Statistical tests, such as t-tests and ANOVA, indicated that USES outperformed traditional measures as it was higher in performance (higher score) and less variable (lower std deviation). This holistic approach ensured that USES captured not only the surface-level overlap but also the deeper semantic meaning and coherence, leading to more accurate and fair assessments of summary quality.

5.3 Discussion of Research Question Second:

What other evaluation techniques can look at the meaning of summaries (semantic) and their coherence beyond simple word overlap measurements?

Summary's differences in wording and structure are the topmost challenge which refers to the phrasing and structure of the reference. The USES provides an effective metric to evaluate summarizing frameworks. USES can more than measuring only word overlap and occurrence which generally other metrics do. USES seems to have better understanding of the meaning and coherence of a summary and it does thorough the various of the other metrics as follows:

1) **BERTScore** takes BERT embeddings as a metric to measure the word level semantic similarity of the generated summary and the reference. Despite accounting meanings in

context, overall coherence and fluency of the summary was not measured through it. By adding BERTScore, USES was a powerful step forwards semantic understanding.

2) **Wu-Palmer Similarity** defines semantic similarity with the help of taxonomic semantic relations. USES assessed summaries formed by the use of synonyms or paraphrases which ensured the meaning of the original text was captured and not just the words.
3) **Cosine Similarity** measured cos distance between the words in any given corpus of the text. USES used this to check overall meaning and presence of the word by arranging word in vector form. Even if the wording in the text differs from the reference, the overall it tries to capture the relatedness of words by looking at vector space.
4) **Sentence Transformer Embeddings** create dense vector representations of a full sentence which help assess the semantic similarity of summary. This helped USES to evaluate whether the summary retains the main ideas and coherence, even with different sentence structures.
5) **ROUGE-1 and METEOR** are metrics that only account for surface level overlap and are not good at semantic similarities.

USES combined these metrics but gave more weightage to the semantic and coherence evaluation lacking in traditional metrics. By using a composite approach of ROUGE-1 and BERTScore as well traditional cosine similarity with sentence transformer embeddings, USES presented more holistic coverage how the generated summary can be evaluated for its true potential.

5.4 Discussion of Research Question Three:

What factors should guide the selection of a set of optimal text summarization evaluation metrics in various summarization experiments?

The best text summarization evaluation metric will depend on what the experiment intends to do, as well as the type of the summary being assessed. Important aspects to take into account while selecting evaluation metrics with focus on USES:

1. When it came to semantic accuracy, we had to see to what extent does the summary retains the meaning of the original content. BERTScore, Wu-Palmer similarity and sentence transformer embeddings were all metrics used that help retain the semantic meaning if there was a use of synonyms, paraphrasing or restructuring.
2. The summary must cover key aspects of the source material. ROUGE and BLEU metrics were used to determine the overlap word. USES used certain semantic similarity metrics, that is, Cosine Similarity, Sentence transformers, and more, to estimate how much summary was overlapping with the original content.
3. A decent summary should be coherent and fluid in nature. BERTScore with sentence transformers can check the semantic coherence of sentences. Also, USES used traditional cosine similarity which further checks the overall coherence of the summary.
4. One has to be adaptable because different types of summarizations require different evaluations. USES can be applied to various types of texts. It combined semantic assessment with surface matching metrics. Therefore, it can be used for different summarizing tasks.
5. Domain-specific factors are being taken into consideration which include the area in which the evaluation is being used. One example of the effect of genre is the presence of scientific or technical detail versus narrative structure common to creative writing.
6. Balance between Precision and Recall: It is essential not to put too much focus on precision; meaning the correctness of the words, but also on recall; meaning coverage of important points. BLEU and ROUGE measured this balance, but USES improved

on this by adding in semantic matching (via BERTScore, WUP and sentence transformers) to provide a more balanced and comprehensive measure.

USES is a hybrid metric that combines many evaluation metrics. These evaluation metrics are BERTScore, Wu-Palmer Similarity, Cosine Similarity, Sentence Transformer embedding, Traditional ROUGE and METEOR. Hence making it very useful for summarization tasks. This helped USES to evaluate the content of the summaries and its structure too. It addressed the issue of single model metrics like BLEU, ROUGE, BERTScore, METEOR. This method examined the different aspects of the summaries and language and text.

Chapter 6 SUMMARY, IMPLICATIONS AND RECOMMENDATIONS

6.1 Summary

In this chapter, we summarize the overall findings/contributions of the study with the implications about USES included in the future research recommendations and the conclusion as well. To overcome the limitations of the commonly used evaluation metrics in the text summarization task, USES was developed. Existing metrics work on word overlap counts which do not capture the full semantic and lexical qualities of a summary.

In the implications section, we discuss how USES improved evaluation of summaries created by Large Language Models. USES gives a better measure of assessing the summaries by combining traditional measure such as ROUGE with more sophisticated measures such as BERTScore, Wu-Palmer similarity and sentence embeddings. This gives rise to enhance

evaluation and fairness in automatic summarization evaluation where ‘paraphrase structure’ matters. Implication on summary quality would be discussed.

In the light of current limitations of USES, particularly its computational complexity issue and relying on a limited set of lexical resources like WordNet, suggestions for future research are made. As discussed in this chapter we can extend the use of USES by:

- Making algorithm more efficient using Fuzzy matching.
- Using domain-specific lexicons.
- Developing light-weight models.

This will make USES scalable and applicable to many domains and languages. The conclusion section summarizes the thesis and its contribution to the field of text summarization evaluation. USES is indeed a valuable framework. It fills a gap that we had before now. USES employs both lexical and deep semantic evaluation techniques to provide a better assessment of the quality of created summaries. Even with a few computational challenges, this brings a good basis for any further developments in the more appropriate evaluation metrics for summarization tasks.

6.2 Real-world business cases for USES

The USES metric for evaluating LLM-generated summaries has diverse use cases across industries, addressing specific needs and improving the quality of generated content.

In **the news and media sector**, where the volume of digital content is rapidly increasing, the metric can ensure that summaries produced by LLMs are factually accurate, retain the critical essence of the content, and meet the required standards.

In a similar way, automated **customer support systems** can use USES to evaluate the quality of summaries created for user queries, user complaints, user resolutions, etc., and later on these summaries will be turned into actionable insights leading to enhanced user experience.

In analyzing **legal documents**, mistakes in summaries can have serious consequences. USES can evaluate the summaries for completeness and correctness. This prevents important information from comprehensive contracts or case documents from being missed; saving time and money. For **e-learning platforms**, the metric USES can ensure that summaries of textbooks, research articles or video lectures are informative, concise. In **social media analytics**, the USES can help businesses assess whether LLM generated summaries of trends, user opinions, or reviews accurately reflect the underlying sentiment and context, enabling informed decisions in marketing and public relations strategies.

6.3 Business Benefits

The new metric USES offers many affordable business benefits. It ensures better summaries in various fields and thus greatly helps in operations, risk, and resource management. Below are key business-oriented benefits.

Reduction in manual review: The proposed metric USES reduces the requirement for human oversight in quality checks by allowing LLMs improved evaluation of summaries. This lowers the cost of operations in sectors like healthcare, law, and finance where manual reviews are done to check for inaccuracies and non-completeness.

Less cost in model development: USES reduces the need to repetitively refine and retrain the model by providing a more realistic and fine-grained mechanism for summary evaluation. By lowering the number of iterations needed to reach a satisfactory level of quality, compared with the use of the BLEU or ROUGE, the cost of LLM development and deployment is lowered as well.

Lowering of regulatory and legal risks: In industries with high stakes, inaccurate or incomplete summaries can lead to compliance breaches, legal penalties, and loss of money. The new measurement USES will help improve evaluation capabilities of LLMs to ensure

summaries meet high accuracy and completeness standards, thus avoiding lawsuits, penalties or costly mistakes in critical sectors.

Enhanced automated workflows: This is one of the more prominent cases of automations. The metric USES is designed to help LLMs enhance their summaries with lesser mistakes thereby helping in streamlining the workflows in customer support, marketing and social media analytics. Businesses can redirect resources and save time and money when the summaries are created correctly in an automated manner.

Quick and faster decision: By generating summaries that are more actionable and in line with business goals, the metric proposed USES reduces the need for any time-consuming validation by decision makers. These not only speed up decision making but also lower the labor costs associated with them. Using the metric will give you useful summaries which reduces the time spent on validations or edits by the decision makers.

6.4 Implication

The Unified Summary Evaluation Score (USES) is a new metric for evaluating text summaries. USES brings together simple metrics at the surface level with more complex metrics at the semantic level. Thus, USES can be useful to assess new summarization models while they are being built and also for assessing existing summarization models.

6.4.1 Improved precision for summarization evaluation

The utilization of USES will increase the accuracy of the evaluation. Standard measures such as BLEU, ROUGE and so on focus on word overlaps by penalizing those summaries that paraphrase or reword the content but do not alter the meaning. USES tackles this through semantic similarity measures such as Wu-Palmer Similarity (WUP) and cosine similarity with sentence embeddings. Using this:

- It can evaluate meaning retention in paraphrased content.

- To offer a more reliable evaluation of summaries that create a more abstract representation rather than a copy of the reference.

Greatly improved near human evaluation in which bigger language model summaries are judged basically on how well they convey the meaning of the original rather than how they fit tighter within a pretty narrow span of reference.

6.4.2 Encouraging more human-like summarization models

The USES metric encourages creating summary models based on meaningful content, not just similarity in words. Since traditional metrics often encourage models to replicate surfaces on other pieces of text, which leads to the replicating of those features, USES on other hand promotes to produce much more relevant and useful outputs. USES will:

- Encourage natural language use, and reward models for capturing meaning while allowing for flexibility in expression.
- An emphasis on fluency and coherence so that the models create summaries that capture key information without making it sound too robot-like.

This trend can spark advancement in model building which can build summarizers that create summaries more like human and are more readable and logically consistent.

6.4.3 Improved utility in real-world applications

The dual ability of USES to assess correctness of content and its semantic consistency makes summarization systems much more useful in real-world contexts. How in-depth assessment could improve the reliability of automated summaries like:

- The accuracy, coherence and brevity of summaries are important in journalism and media.
- Healthcare and legal services are two areas where precision and clarity matter for summarizing a lot of complex information.

Because of this, professional quality summaries can be trusted for better decisions and faster processing of time-critical information.

6.4.4 Advancing research in text summarization

By enabling a more comprehensive assessment framework, USES could advance the field of text summarization research. Its surface and semantic metrics allows researchers to:

- Benchmark models more accurately in terms of their strengths and weaknesses.
- We encourage the development of new model architectures that emphasize improving semantic coherence and fluency, rather than simply optimizing for word overlap.

So, USES can invigorate the development of next-generation summarization models and thereby foster innovation and more meaningful progress in NLP research.

6.4.5 Better alignment with human judgment

Another important implication of USES is the ability to be more aligned to human evaluation criteria. Usually, human summarized assessments focus on semantic understanding, fluency and coherence. However, traditional metrics do not account for these three factors. The solution offered by USES is:

- To include semantic similarity measures which reflect human judgment of the meaning.
- Evaluating both sentence-level coherence and overall text fluency, ensuring that summaries are logically structured and easy to follow.

When automated assessment systems fit human judgement, they become more agreeable. They help to strengthen the reliability and authenticity of better AI summaries.

6.4.6 Increased adaptability across domains

With its ability to combine both traditional and semantic evaluation metrics, USES can be used in different domains and summarization tasks. For instance:

- Scientific papers have a high demand for accuracy and content coverage. For such content, the automatic evaluation metric has been used which is ROUGE and METEOR. But along with that, an explanation of semantic fidelity is also necessary which is made around BERTScore and sentence embeddings.
- The usability of USES for creative writing and narrative summaries is to assess their coherence and fluency. Moreover, it rewards paraphrasing and stylistic adaptations.

USES is adaptable, making it valuable in multi-domain applications where evaluating different summary goals (for example, technical accuracy vs. narrative flow) requires different criterias.

6.4.7 More comprehensive evaluation framework

The USES metric takes a holistic approach that combines traditional metrics with advanced techniques into one overall framework to evaluate arguments or evidence. This metric can evaluate summaries from different perspectives, which includes:

- Semantic accuracy (BERTScore, WUP, Cosine Similarity)
- Surface-level precision and recall (ROUGE, METEOR) and
- Fluency and Coherence (using sentence transformers)

This study enlists the testing of multiple metrics with different summaries and languages. This type of evaluation ensures that the summary is not only factually correct but also conveys the meaning intended and is logically structured.

6.4.8 Reducing bias towards word matching

Benchmark metrics such as BLEU and ROUGE frequently impose excessive penalties on summaries that do not match the reference text even in instances where meaning is preserved. The stress on semantic similarity in USES reduces this bias and allows for an unbiased evaluation of summaries with synonyms, paraphrases or differently worded sentences that convey the same meaning.

As a result less biased evaluations arise. This is especially important for evaluation of the quality of abstractive summaries as these differ in wordings.

6.4.9 Enhancing multi-task and multi-document summary evaluations

USES is good for exploring multi document summarizations due to its adaptable semantics. Standard metrics cannot evaluate summaries that do not use the same words as the reference especially when in different documents. By emphasizing meaning and coherence, USES can help to:

- Better evaluate multi-document summary, i.e., summaries from different documents.
- Assist learning models that are multi tasking i.e., various approaches for summarization (extractive, abstractive) need different evaluation strategies.

The flexibility of USES in NLP tasks will enhance the usability in various industries.

6.5 Recommendations for Future Research

Although USES is a leap forward in evaluation metrics, it poses limitations and there exist several avenues for future work in areas of computational efficiency, scalability and lexical similarity. To effectively address these challenges, it would be critical to optimize USES and making it flexible and broadly applicable to many summarization tasks.

6.5.1 Enhancing Computational Efficiency and Speed

A major limiting factor of USES is its computational intensity, which arises from its deployment of several metrics, each of which is resource-intensive. Future research should concentrate on:

- **Algorithm Optimization**: Approximate existing algorithms to run faster without significantly compromising the quality of evaluation. Utilization of dimensionality reduction or embedding quantization can reduce the computation.

- **Hardware Acceleration**: Using GPU or TPU based acceleration that also applies parallel processing to the task of Deep Learning can dramatically lower run time without affecting accuracy.
- **Consolidating Metrics**: One way to boost the speed of performance metrics is to consolidate them. For instance, BERTScore and sentence embeddings are both semantic metrics. Perhaps there's a way to combine them into one metric. That would save on computing time and improve the performance metrics overall.

6.5.2 Expanding Lexical Databases Beyond WordNet

USES relies on WordNet for measuring lexical similarity, which might be a limitation in some domains. Future research can build on the below:

- By Integrating domain specific lexicons: By adding these lexicons (e.g., UMLS for medical texts or legal dictionaries), one may obtain better performance in technical domains.
- Leveraging modern lexical resources: The USES metric can be improved in two ways. One is by exploiting modern lexical resources like ConceptNet or FrameNet or similar large-scale pre-trained semantic resources. By doing this, the USES system might be able to cover more lexical relations.
- Contextual embeddings: Using new context-aware models like T5 or GPT-based embedding may provide complete, rich evaluations of words with flexible, convenient, intra-lingual and cross-lingual affordances.

6.5.3 Developing Lightweight Semantic Models

USES, which depends on heavy models like BERT and sentence transformers, is costly on computation. Future explorations can be done on the below:

- Creating light-weight semantic models through knowledge distillation and similar methodologies which results in a smaller and efficient model without much loss in accuracy.
- Fine-tuning models for summarization tasks will lessen the computing needs along with maintaining evaluation quality.

6.5.4 Expanding Multilingual and Cross-Lingual Capabilities

To help with summary evaluation in multiple languages, future research should look at how well USES deals with non-English summaries:

- Multilingual embeddings - This involves using newer multilingual embeddings (e.g., mBERT or XLM-R) that will allow USES to fairly evaluate summaries in other languages.
- Including lexical resources like BabelNet and EuroWordNet may improve the accuracy of lexical similarity measures in cross-lingual contexts.
- Employing machine translation as part of the evaluation pipeline may be useful because misalignment means that even summaries with the same core meaning may not match in the overlap.

6.5.5 Enhancing Coherence and Fluency Metrics

Although USES combines evaluations of sentence-level coherence and fluency, global coherence and summary's naturalness can be better evaluated. Future research could look into:

- Discourse-level models: These evaluate discourse coherence across an entire summary. This may help ensure the summary is not only semantically accurate but also well-structured and logically organized.
- Improving fluency metrics of sentences used for summary generation may help in ensuring grammaticality, variation in sentence lengths and overall readability. This will enable USES to become a better judge of generated summaries and their naturalness.

6.6 Conclusion

This thesis fills a void in measuring text summarization by proposing the Unified Summary Evaluation Score (USES), a new evaluation metric that incorporates traditional as well as contemporary semantic evaluation metrics. Contemporary summaries are produced using methods which may statistic, linguistic and transformer based models. The summary produced should be able to capture both lexical and semantic meaning. However, this was not the case with all the automatic evaluation metrics used. Commonly used measures such as ROUGE and BLEU stress on surface-level word overlaps and do not adequately capture deeper semantic aspects which lead to incomplete or biased evaluations.

The gap identified in the literature suggests that most of the work on summarization evaluations still relies heavily on metrics that focus on lexical overlap and ignore how well the summary captures the meaning/structure of the text. This leads to an evaluation bias especially for abstractive summaries that paraphrase or reorder while being accurate. In light of this, it became important to have a more elaborate evaluation framework that can capture the lexical and semantic components.

To fill this gap, USES was developed. It combines standard measures like **ROUGE-1** and **METEOR** with advanced techniques like Wu-Palmer similarity, BERTScore, Cosine Similarity and sentence transformer embedding. Addition of USES to the evaluation measure allows to go beyond word overlap and evaluate summaries for their meaning, coherence at the sentence level and fluency. Using many models together gives you the best assessment of Large Language Model outputs. This is particularly true for abstractive summary tasks.

While USES is truly a step forward, it does bring in limitation with respect to excessive time required for the process. Combining different metrics that need more and more resources is an expensive affair. Thus, USES is recommended for offline summarization. Moreover, while the

tool's use of WordNet for calculating lexical similarity is effective in itself, it may benefit from more diverse and domain specific lexical resources.

In conclusion, the thesis proposes USES that improves the existing state of the art text summarization evaluation which has been mostly ignored by research effort so far. Furthermore, USES overcomes the drawback of traditional metrics by offering a holistic, semantic-based evaluation of the summary. USES aims at correcting the deficiencies in existing evaluation practices to allow for more accurate and context-sensitive to assess summaries that would assist in the continuous development of advanced summarization models. Proposed novel metric USES is computationally cheap and generic, but there is a need for further research and optimization specifically designed for a particular domain.

BIBLIOGRAPHY

Banerjee, S. and Lavie, A. (2005) 'METEOR: An automatic metric for MT evaluation with improved correlation with human judgments', *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, pp. 65–72.

Bordia, S. and Bowman, S.R. (2019) 'Identifying and reducing gender bias in word-level language models', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, Hong Kong, pp. 1–10.

ChatGPT (2023) 'ChatGPT: Optimizing language models for dialogue', *OpenAI*. Available at: https://openai.com/chatgpt (Accessed: 11 November 2024).

Chen, Y., Hovy, E. and Tang, D. (2016) 'A comprehensive analysis of the challenges in summarizing multi-document news articles', *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, Texas, pp. 1583–1592.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, New Orleans, USA, pp. 4171–4186.

Humby, C. (2006) *Data is the new fuel*.

Howard, J. and Ruder, S. (2018) 'Universal language model fine-tuning for text classification', *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339.

Hermann, K.M., Blunsom, P., Kayser, S., Liao, S. and Watson, P. (2015) 'Teaching machines to read and comprehend', *Proceedings of the 2015 Conference on Neural Information Processing Systems (NeurIPS 2015)*, Montreal, Canada, pp. 1693–1701.

Kallimani, P., Srinivasa, K. and Eswara Reddy, P. (2016) 'A unified model for document condensing and key information extraction using attribute-based information extraction rules and class templates', *International Journal of Computer Applications*, 139(5), pp. 26–30.

Kryściński, W., McCann, B., Xiong, C. and Socher, R. (2019) 'Evaluating the factual consistency of abstractive text summarization', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, pp. 2901–2911.

Le, T. and Le, D. (2013) 'Abstractive text summarization using discourse rules and syntactic constraints', *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, pp. 140–145.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Khandelwal, U., Schuster, M., Chen, D., Riley, P., Liu, X. and Parikh, A. (2019) 'BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension', *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, Long Beach, California, pp. 6905–6916.

Lin, C.Y. (2004) 'ROUGE: A package for automatic evaluation of summaries', *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, pp. 25–26.

Mehdad, Y., Carenini, G. and Ng, R.T. (2014) 'Abstractive summarization of spoken and written conversations based on phrasal queries', *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 1220–1230.

Nallapati, R., Zhai, F. and Zhou, B. (2016) 'Abstractive text summarization using attentional

encoder-decoder recurrent neural networks', *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, San Diego, USA, pp. 11–19.

Oya, T., Takahashi, S., Seki, Y. and Imai, M. (2014) 'Automatic abstractive summarization of meeting conversations using a multi-sentence fusion technique', *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, Singapore, pp. 3101–3105.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) 'BLEU: A method for automatic evaluation of machine translation', *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, USA, pp. 311–318.

Pimpalshende, N. and Mahajan, A. (2016) 'A rule-based approach for summarizing historical documents', *International Journal of Computer Applications*, 136(6), pp. 22–26.

Raffel, C., Shazeer, N., Roberts, A., Lee, S., Narang, S., Matena, M., Zhou, Y., Faydice, A., Liu, P.J. and Smith, N.L. (2019) 'Exploring the limits of transfer learning with a unified text-to-text transformer', *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, pp. 1–16.

Rush, A.M., Chopra, S. and Weston, J. (2015) 'A neural attention model for abstractive sentence summarization', *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisbon, Portugal, pp. 379–389.

See, A., Liu, P.J. and Manning, C.D. (2017) 'Get to the point: Summarization with pointer-generator networks', *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. And Polosukhin, I. (2017) 'Attention is all you need', *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, California, pp. 5998–6008.

Vinyals, O. and Le, Q. (2015) 'A neural conversational model', *Proceedings of the 32nd International Conference on Machine Learning*. Available at: https://arxiv.org/abs/1506.05869 (Accessed: 11 November 2024).

Vodolazova, D. and Lloret, J. (2019) 'Enhancing abstractive text summarization through syntactic text simplification and subject-verb-object concept frequency scoring', *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2019)*, pp. 1075–1080.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. (2019) 'GLUE: A

multi-task benchmark and analysis platform for natural language understanding', *Proceedings of the International Conference on Learning Representations (ICLR)*. Available at: https://gluebenchmark.com (Accessed: 11 November 2024).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Guo, H. and Rush, A.M. (2020) 'Transformers: State-of-the-art natural language processing', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 38–45. Available at: https://arxiv.org/abs/1910.03771 (Accessed: 11 November 2024).

Yahya Saeed, M., Awais, M., Younas, M., Arif Shah, M., Khan, A., Irfan Uddin, M. and Mahmoud, M. (2021) 'An abstractive summarization technique with variable length keywords as per document diversity', *Computers, Materials & Continua*, 66(3), pp. 2409–2423. doi: https://doi.org/10.32604/cmc.2021.014330.

Zhang, J., Yasunaga, M., Chen, P., Li, J., Liu, P.J. and Lee, L. (2020) 'PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Virtual, pp. 1–7.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M. and Eger, S. (2019) 'MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance', *arXiv.org*. Available at: https://arxiv.org/abs/1909.02622 (Accessed: 11 November 2024).