NLP TRANSFORMERS: ANALYSIS OF LLMS AND TRADITIONAL APPROACHES FOR ENHANCED TEXT SUMMARIZATION

Yunus Emre ISIKDEMIR[1*]
[1] Computer Engineering Department, Hacettepe University, Ankara, Turkey,
ORCID No : https://orcid.org/0000-0001-7022-2854

Keywords
Text Summarization
Transformers
NLP
LLM
Deep Learning

Abstract

As the amount of the available information continues to grow, finding the relevant information has become increasingly challenging. As a solution, text summarization has emerged as a vital method for extracting essential information from lengthy documents. There are various techniques available for filtering documents and extracting the pertinent information. In this study, a comparative analysis is conducted to evaluate traditional approaches and state-of-the-art methods on the BBC News and CNN/DailyMail datasets. This study offers valuable insights for researchers to advance their research and helps practitioners in selecting the most suitable techniques for their specific use cases.

TRANSFORMER MİMARİSİ TABANLI METİN ÖZETLEME VE DOĞAL DİL ANLAMLANDIRMA

Anahtar Kelimeler
Metin Özetleme.
Transformer
Doğal Dil İşleme
Büyük Dil Modelleri
Derin Öğrenme

Öz

Hızla büyüyen yüksek hacimli bilgi kaynaklarından faydalı bilgilere erişim, giderek zorlaşmaktadır. Bu soruna çözüm olarak geliştirilen metin özetleme yöntemleri, yüksek hacimli belgelerden önemli bilgilerin çıkarılmasında önemli bir role sahiptir. Belgeleri filtreleme ve ilgili bilgileri çıkarma amacıyla çeşitli teknikler mevcuttur. Bu çalışmada, BBC News ve CNN/DailyMail veri setleri üzerinde, geleneksel yaklaşımlar ile en güncel yöntemlerin karşılaştırmalı analizini sunmaktadır. Araştırmacılara ilerlemelerine katkı sağlayacak değerli bilgiler sunmakta ve uygulayıcıların özel kullanım durumlarına en uygun teknikleri seçmelerine yardımcı olmaktadır.

Araştırma Makalesi		Research Article	
Başvuru Tarihi	: 27.05.2023	Submission Date	: 27.05.2023
Kabul Tarihi	: 12.03.2024	Accepted Date	: 12.03.2024

[*] Sorumlu yazar: yemreisikdemir@gmail.com
https://doi.org/10.31796/ogummf.1303569

1. Introduction

In the age of excessive information, the ability to distill key insights from vast amounts of text data has become increasingly crucial. Text summarization, a subfield of natural language processing (NLP), has emerged as a valuable solution to tackle this challenge. By automatically condensing lengthy documents into concise summaries, text summarization techniques enable individuals to extract relevant information efficiently and make informed decisions.

A significant advancement in the field of NLP has been the introduction of transformer-based models. These models have revolutionized the way we approach various natural language understanding tasks. Generative Pre-trained Transformer (GPT) is an example of such models. They leverage large-scale neural networks with multiple layers. The purpose is to capture intricate linguistic patterns and generate coherent and contextually rich outputs. These language models, trained on extensive amounts of text data, have demonstrated remarkable capabilities in tasks such as machine translation, sentiment analysis, and question answering.

The emergence of large language models has had a profound impact on numerous domains, ranging from content creation and customer support to legal research and healthcare. With their ability to comprehend and generate human-like text, these models have the potential to enhance productivity, improve decision-making processes, and unlock novel applications in various industries. Text summarization stands as a significant beneficiary of these advancements, as large language models offer powerful tools for automatically generating accurate and informative summaries from diverse textual sources.

This research paper aims to delve into the importance of text summarization, NLP transformers, and large language models in the context of information retrieval and comprehension. By exploring the underlying mechanisms of text summarization algorithms, examining the capabilities of transformer-based models, and discussing real-world applications, this study seeks to shed light on the potential benefits and challenges associated with leveraging these technologies.

In the following sections, it will be discussed the fundamental concepts of text summarization, highlighting different approaches and techniques employed to extract salient information. Subsequently, we will delve into the architecture and working principles of NLP transformers, elucidating their role in facilitating advanced natural language processing tasks. Finally, we will examine the impact of large language models on text summarization, emphasizing their significance in empowering users to make effective use of textual data by comparative analysis.

By comprehensively analyzing the intersection of text summarization, NLP transformers, and large language models, this research aims to contribute to the growing body of knowledge in this field, providing insights that can inform both researchers and practitioners in leveraging these technologies for information extraction and comprehension.

2. Related Works

The explosive growth of digital data in recent years has made it increasingly difficult to find relevant information among the vast amount of available data. To address this challenge, automatic text summarization has emerged as an effective solution that generates a concise summary of a large document containing essential information. In recent years, advancements in NLP and machine learning have greatly improved text summarization algorithms. State-of-the-art models can now produce informative and coherent summaries that are also easy to read.

There are two main categories of text summarization that are extractive and abstractive (El-Kassas, Salama, Rafea and Mohamed, 2021). Extractive summarization involves selecting important sentences or phrases from a document and combining them into a summary (Liu, Chen and Chen, 2020), while abstractive summarization generates a new summary by synthesizing information from the document (Zheng et al., 2020). Extractive summarization techniques are generally considered easier to implement and have been extensively studied. On the other hand, abstractive summarization is more challenging, requiring the system to understand the semantics of the input document and generate novel sentences (Gupta and Gupta, 2019).

Academic research in the field of text summarization is increasing, as this powerful tool gains significant attention for its diverse range of practical applications. One of the primary areas of focus in this research is the development of more advanced algorithms and techniques for generating accurate and informative summaries of various types of text, including news articles, legal documents, and scientific papers (Altmami and Menai, 2022).

Rodríguez-Vidal et al. (2011) presents an introduce a novel test compilation comprising reputation reports, which condense tweet streams pertaining to 31 companies operating in the banking and automobile sectors. Distinguishing itself from the conventional summarization task, the generation of reputation reports necessitates the inclusion of priority signals in order to effectively tackle the task at hand. This methodology leverages an analogy between reputation reports and the issue of diversity in search, resulting in tangible benefits.

Widjanarko, Kusumaningrum and Surarso (2018) proposed a solution to help people acquire comprehensive online news information in Indonesian by using Latent Dirichlet Allocation (LDA) for automatic multi-document summarization based on sentence extraction. They employed an unsupervised learning approach and evaluated the performance of LDA and Significance Sentence methods in calculating sentence weights, using a range of alpha and beta parameters and compression rates.

Goularte, Nassar, Fileto and Saggion (2019) in their study, propose an automatic process for text assessment using fuzzy rules and correlated features to summarize texts with a relatively small number of fuzzy rules. They also conducted a comparative analysis of their proposed method against various existing approaches, and demonstrated its effectiveness by achieving promising results in terms of f-measure.

In their work, Cagliero, Garza, and Baralis (2019) present a novel approach to multilingual summarization. Their proposed algorithm utilizes frequent itemset mining and Latent Semantic Analysis (LSA) to extract concise summaries from sets of textual documents. The algorithm employs a combination of itemset and LSA-based strategies, using a greedy approach to select sentences that effectively cover latent concepts with minimal redundancy.

Widyassari et al. (2022) conducted a thorough review of text summarization studies from 2008 to 2019. Their investigation encompassed 85 scholarly articles from journals and conferences. The aim was to elucidate the prevailing research themes, emerging trends, as well as the datasets and preprocessing methods commonly employed in this domain of research. However, they did not address or discuss the latest NLP Transformers in their analysis.

Manojkumar, Mathi and Gao (2023) aims to find the most appropriate algorithm to summarize food or restaurant reviews and enable users to save time while comprehending the reviews' essence. To achieve this goal, the authors conducted an empirical analysis of various unsupervised extraction-based text summarization techniques and determined that the LexRank method shows promising results.

Liu and Lapata (2019) introduce a general framework for text summarization using Bidirectional Encoder Representations from Transformers (BERT) as a document-level encoder. Their framework includes both extractive and abstractive models, with an extractive model built on top of the document-level encoder using intersentence Transformer layers. They also propose a new fine-tuning schedule for abstractive summarization that uses different optimizers for the encoder and the decoder to address the mismatch between the pretrained encoder and the untrained decoder.

Zhang, Wei and Zhou (2019) address the challenges of training neural extractive summarization models using heuristically created sentence-level labels. They propose a method called HIBERT (Hierarchical Bidirectional Encoder Representations from Transformers) for document encoding, which uses a pre-trained hierarchical transformer on unlabeled data. The pre-trained HIBERT is applied to the summarization model, resulting in significant improvements in performance on the CNN/DailyMail and New York Times datasets.

Lewis et al. (2019) presents an autoencoder architecture called Bidirectional Auto-Regressive Transformers (BART) that learns to reconstruct the original text from a corrupted version created by an arbitrary noising function. The model is based on a Transformer architecture and generalizes previous pre-training schemes such as BERT and GPT. The researchers assess various methods of introducing noise and determine that the most effective performance is achieved through the random rearrangement of sentence order and the utilization of a novel technique for filling in the gaps.

Zhang, Zhao, Saleh and Liu (2020) propose a pre-training objective for large Transformer-based encoder-decoder models specifically tailored for abstractive text summarization. A novel approach known as PEGASUS is introduced, which includes the extraction or concealment of significant sentences in an input document. These sentences are then combined into a single output sequence alongside the remaining sentences, resembling an extractive summary. The authors assess the performance of their most effective PEGASUS model on 12 distinct summarization tasks across various domains. Their evaluation shows that the model achieves state-of-the-art results on all 12 datasets, as measured by ROUGE scores. They also show that their model is able to adapt to low-resource summarization datasets very quickly and achieves human-like performance on multiple datasets using human evaluation.

The following sections provide a detailed examination of the essential elements related to text summarization. This study thoroughly investigates a range of crucial components, including various text summarization techniques, modeling approaches, evaluation metrics, and preprocessing techniques.

3. Background

In this section, our focus is to present the terminologies related to text summarization. The purpose is to enhance the reader's understanding of the remaining sections of the paper, as well as the experimental outcomes.

3.1. Text Summarization Techniques

There are two primary methods of achieving text summarization, extractive and abstractive, which are selected based on the specific problem at hand. In this section, each method will be investigated in detail.

3.1.1. Extractive Summarization

Extractive summarization is a subfield of NLP that involves automatically identifying and extracting the most important information from a text document and presenting it in a condensed form. The extracted information is usually presented in the form of a summary, which captures the main points of the original text and allows users to quickly understand the content of the document without having to read the entire thing. This method, only selects and presents information that is already present in the text instead of generation.

There are several techniques used in extractive summarization. One common approach is to use sentence scoring, which involves assigning a score to each sentence in the document based on its relevance to the overall content (Verma and Om, 2019). The sentences with the highest scores are then selected to be included in the summary shown in Figure 1.

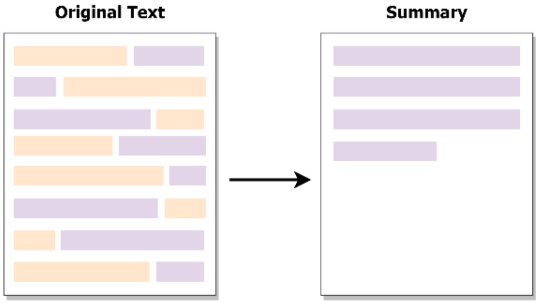

Figure 1. Extractive Text Summarization

Sentence scoring can be done using various methods, such as:

Frequency-based methods: These methods assign a score to each sentence based on the frequency of the words it contains. Sentences that contain more frequently occurring words are considered more important (Abdel-Salam and Rafea, 2022) shown in Figure 2.

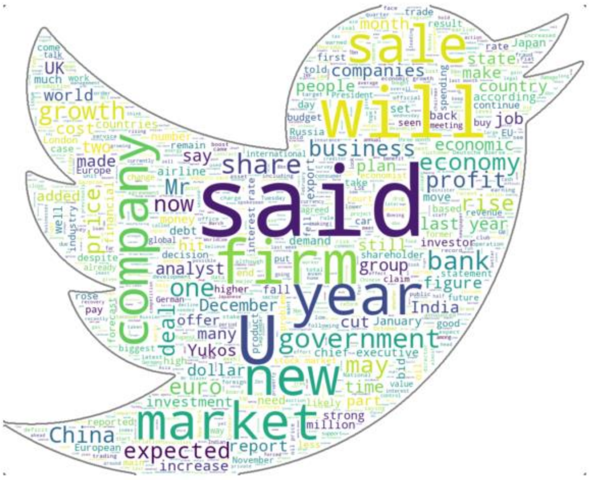

Figure 2. Most Frequent Words in Entertainment Category of BBC Dataset

Graph-based methods: These methods represent the document as a graph (Vhatkar, Bhattacharyya and Arya, 2020), where the sentences are nodes and the edges between them represent their semantic similarity. The sentences with the highest centrality in the graph are considered the most important shown in Figure 3.

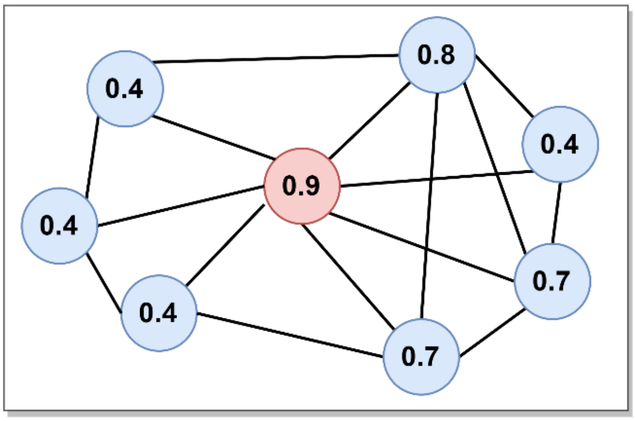

Figure 3. Most Frequent Words in Entertainment Category of BBC Dataset

Machine learning-based methods: These methods involve training a machine learning model to predict the importance of each sentence (Abdelaleem, Kader and Salem, 2019) based on various features such as the length of the sentence, the presence of certain keywords, and the context in which the sentence appears.

3.1.2. Abstractive Summarization

Abstractive summarization is a NLP technique that involves generating a summary of a text document that is not a verbatim copy of any sentence in the original document. Instead, the summary is generated by synthesizing the important information from the document in a way that captures the essence of the original text.

Unlike extractive summarization, which simply selects the most important sentences from a document, abstractive summarization involves understanding the meaning of the text and generating new sentences to express that meaning in a more concise form (Ramina, Darnay, Lubde and Dhruv, 2020).

The process of abstractive summarization involves following steps:

- **Text pre-processing:** The text is pre-processed to remove noise and irrelevant information, such as stop words and punctuation.
- **Text representation:** The text is represented in a way that can be easily processed by a machine learning model. This could involve techniques such as tokenization, stemming, and vectorization.
- **Neural network training:** A neural network is trained to learn the relationship between the input text and the desired summary. The neural network could use techniques such as encoder-decoder models, attention mechanisms, and reinforcement learning to generate a summary (Syed, Gaol and Matsuo, 2021).
- **Summary generation:** Once the neural network is trained, it can be used to generate a summary for a given input text. The generated summary should capture the essence of the original text in a concise and coherent form.

There are several challenges associated with abstractive summarization. One major challenge is the ability of the neural network to capture the nuances and context of the original text. This requires a deep understanding of the meaning and structure of language, as well as the ability to generate coherent and grammatically correct sentences. Another challenge is the ability of the neural network to deal with out-of-vocabulary (OOV) words (Shi, Keneshloo, Ramakrishnan and Reddy, 2021), which are words that are not present in the training data. This can be addressed by using techniques such as word embedding and transfer learning.

Abstractive summarization has many applications, such as summarizing news articles, social media posts, and product reviews. However, it is still an active area of research, and there are many open challenges and opportunities for improvement.

3.2. Modeling Approaches

Numerous approaches to text summarization exist in the literature, each utilizing distinct techniques and perspectives to address the challenge of shortening text while preserving its key content. This section aims to explore these approaches, as text summarization is a crucial task in NLP that can enhance information retrieval and consumption. Various modeling approaches, including statistical and machine learning-based techniques, have been proposed for text summarization, with some relying on extractive methods and others on abstractive ones. To determine the most suitable approach for specific applications, it is important to compare the strengths and weaknesses of these approaches and evaluate their performance using different metrics.

3.2.1. Unsupervised Learning Approach

Unsupervised learning approaches for text summarization involve training a model on a large corpus of unannotated data (Joshi, Fidalgo, Alegre and Fernandez, 2019), without any human supervision. These models aim to extract the most important information from the input text and generate a summary that captures the key points.

Clustering: In this approach, the input text is first segmented into smaller units, such as sentences or paragraphs. Next, a clustering algorithm is applied to group similar units together. The resulting clusters represent different topics or themes in the input text. Finally, a summary is generated by selecting the most representative units from each cluster (Nazari and Mahdavi, 2019).

Graph-based: The approach is based on graph-based methods. In this approach, the input text is represented as a graph, where nodes represent words or sentences, and edges represent their relationships (Ramesh, Srinivasa and Pramod, 2014). The graph is then analyzed to identify the most important nodes, based on metrics such as degree centrality or betweenness centrality. These important nodes are then used to construct a summary that captures the key information in the input text.

Topic Modeling: In this approach, a topic model is trained on the input text to identify the underlying topics or themes. The most representative sentences for each topic are then selected to generate a summary that captures the key points of the input text (Wu et al., 2017).

3.2.2. Supervised Learning Approach

Supervised learning approaches for text summarization involve training a model on a labeled dataset, where each input text is paired with its corresponding summary. These models aim to learn the mapping between input texts and their corresponding summaries and generate summaries for new texts based on this learned mapping.

Sequence-to-sequence (Seq2Seq) models: In this approach, the input text is first encoded into a fixed-length vector using an encoder network, such as a long short-term memory (LSTM) which is a type of recurrent neural network (RNN) and specifically designed to handle the vanishing gradient problem (Isikdemir and Yavuz, 2022). The encoder network learns to capture the meaning and context of the input text. Next, a decoder network, such as an RNN or a transformer, is used to generate a summary based on the encoded input text (Li, Xu, Li and Gao, 2018). The decoder network learns to generate a summary that captures the most important information in the input text, while minimizing the loss between the generated summary and the ground truth summary.

Pointer-generator networks: In this approach, the model learns to generate summaries by selectively copying words from the input text or generating new words (See, Liu and Manning, 2018). The model consists of an encoder network as shown in Figure 4, which encodes the input text into a fixed-length vector, and a decoder network, which generates the summary by selectively copying words from the input text or generating new words. The pointer-generator network can handle out-of-vocabulary words and generate more fluent summaries compared to Seq2Seq models.

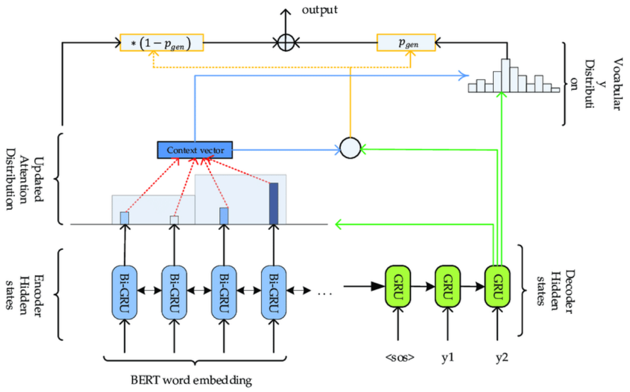

Figure 4. Pointer-generator Network (Wang et al. 2019)

Transformers: In this approach, the input text is first encoded into a sequence of vectors using a transformer encoder network as shown in Figure 5, such as BERT or GPT. The transformer encoder network learns to capture the meaning and context of the input text by leveraging a pre-trained language model that has been trained on a large corpus of texts. Next, a transformer decoder network is used to generate the summary based on the encoded input text (Cai, Shen, Peng, Jiang and Dai, 2019). The transformer decoder network learns to generate a summary that captures the most important information in the input text, while minimizing the loss between the generated summary and the ground truth summary.

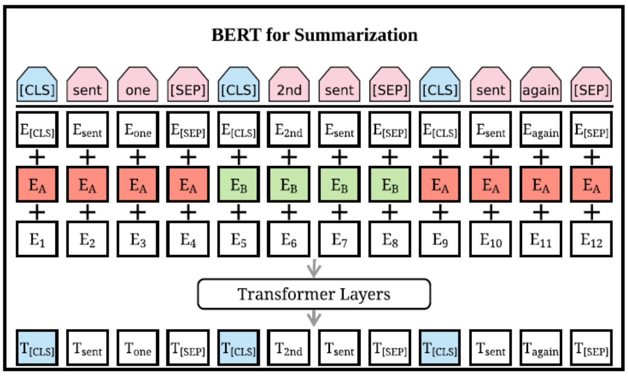
BERT for Summarization
[CLS] sent one [SEP] [CLS] 2nd sent [SEP] [CLS] sent again [SEP]
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ $E_{[CLS]}$ E_{2nd} E_{sent} $E_{[SEP]}$ $E_{[CLS]}$ E_{sent} E_{again} $E_{[SEP]}$
E_A E_A E_A E_A E_B E_B E_B E_B E_A E_A E_A E_A
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12}
Transformer Layers
$T_{[CLS]}$ T_{sent} T_{one} $T_{[SEP]}$ $T_{[CLS]}$ T_{2nd} T_{sent} $T_{[SEP]}$ $T_{[CLS]}$ T_{sent} T_{again} $T_{[SEP]}$

Figure 5. Transformer Architecture (Liu and Lapata, 2019)

Reinforcement Learning: In this approach, the model learns to generate summaries by maximizing a reward signal that measures the quality of the generated summary. The model consists of an encoder network, which encodes the input text into a fixed-length vector, and a decoder network, which generates the summary (Yao, Zhang, Luo and Wu, 2018). During training, the model generates multiple summaries for each input text, and the best summary is selected based on a reward signal that measures the quality of the generated summary.

Supervised learning approaches for text summarization offer advantages such as the production of coherent and readable summaries when compared to unsupervised methods. They excel at handling texts with intricate or domain-specific language, thanks to their training on labeled datasets that encompass the specific language and content of input texts. However, these approaches necessitate abundant labeled data for effective training and might struggle to generalize well beyond the training dataset (Moratanch, 2017). On the other hand, deep learning approaches for text summarization possess their own set of advantages. They are capable of generating high-quality summaries that effectively capture the most crucial information from the input text. Furthermore, they exhibit proficiency in processing texts featuring intricate or domain-specific language, as the models can learn to comprehend the specific language and content of the input texts. Nevertheless, these deep learning methods require substantial amounts of data for training and may involve computational expenses during both training and deployment stages (Yadav, 2022).

3.3. Evaluation Metrics

Evaluating the quality of a text summary is a challenging task, as it involves measuring how well the summary captures the key information and important details of the original text. Several evaluation metrics have been proposed to measure the quality of text summaries, and each metric has its own strengths and weaknesses.

ROUGE: ROUGE (Lin, 2004) stands for Recall-Oriented Understudy for Gisting Evaluation and is a family of metrics that measure the overlap between the generated summary and the reference summary at different levels of granularity, such as word, sentence, or n-gram (Equation (1), (2) and (3)). ROUGE metrics are widely used in the text summarization community because they are simple to compute and correlate well with human judgments of summary quality.

$$ROUGE_1 = \frac{\text{Number of overlapping unigrams}}{\text{Number of unigrams in reference summary}} \tag{1}$$

$$ROUGE_2 = \frac{\text{Number of overlapping bigrams}}{\text{Number of bigrams in reference summary}} \tag{2}$$

$$ROUGE_L = \frac{\text{Longest common subsequence (LCS) length}}{\text{Number of words in reference summary}} \tag{3}$$

METEOR: Meteor, stands for Metric for Evaluation of Translation with Explicit ORdering, is a special metric designed to assess the effectiveness of machine translation and text summarization systems. It employs a weighted F-score based on unigram matches and penalizes deviations in word order (Equation (4), (5), (6) and (7)). The F1 score serves as a metric for assessing both precision and recall in the generated summary compared to the reference summary. This score is calculated by determining the harmonic mean of precision and recall scores (Goutte and Gaussier, 2005). Precision evaluates the proportion of the generated summary found in the reference summary, while recall assesses the percentage of the reference summary captured in the generated summary.

$$P = \frac{\mathrm{m}}{\mathrm{c}}, R = \frac{\mathrm{m}}{\mathrm{r}} \tag{4}$$

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \quad (5)$$

$$Penalty = \gamma \left(\frac{C_m}{\mathrm{m}}\right)^{\beta} \quad (6)$$

$$METEOR = F_{mean} \cdot (1 - Penalty) \quad (7)$$

where C_m represents matching chunks, a chunk being a set of adjacent unigrams in both reference and candidate. m counts mapped unigrams between reference and candidate. The maximum penalty is determined by the value of γ, while the functional relationship between fragmentation and the penalty is influenced by the value of β.

BLEU: Bilingual Evaluation Understudy, BLEU (Papineni, Roukos, Ward and Zhu, 2002), is a popular evaluation metric used in text summarization to measure the quality of machine-generated summaries by comparing them with one or more reference summaries created by humans (Equation (5), (6) and (7)). This metric calculates the similarity between the generated summary and the reference summaries by comparing the words that are common between them. The BLEU score ranges from 0 to 1, where a score of 1 means the generated summary perfectly matches the reference summaries. While BLEU is widely used, it has some limitations, such as not considering the coherence and meaning of the summary, and being biased towards shorter summaries.

$$BLEU = BP \times exp\left(\frac{1}{n}\sum_{i=1}^{n} log(p_i)\right) \quad (5)$$

$$BP = \begin{cases} 1, & if\ c > r \\ \exp\left(1 - \frac{r}{c}\right), & if\ c \le r \end{cases} \quad (6)$$

$$p_i = \frac{\text{Number of matching n-grams}}{\text{Number of n-grams in candidate}} \quad (7)$$

where the variable n represents the maximum order of n-grams used for comparison, c is the length of the candidate translation, and r is the length of the reference translation. BP represents the brevity penalty term, which adjusts the BLEU score to penalize shorter translations.

Intiutive: Human evaluation is also an important component of text summarization evaluation. Human evaluation involves having human judges rate the quality of the generated summaries (Bhandari, Gour, Ashfaq, Liu and Neubig, 2020) based on various criteria, such as readability, coherence, and informativeness. Human evaluation is often used to validate the results of automatic evaluation metrics and to provide a more nuanced understanding of the quality of the generated summaries.

Evaluating the quality of text summarization is a complex task that requires the use of multiple evaluation metrics and human evaluation. ROUGE, F1 score, and BLEU are some of the commonly used automatic evaluation metrics, while human evaluation is used to provide a more nuanced understanding of the quality of the generated summaries.

4. Methodology

The primary aim of this study is to compare and assess the effectiveness of different document summarization techniques on the BBC news (Greene, and Cunningham, 2006) and CNN/DailyMail (Hermann et al., 2015) dataset to provide valuable insights for both researchers and practitioners in the field. The BBC news dataset consists of 2225 documents, obtained from the BBC news website, covering five various topics including business, entertainment, politics, sport, and technology. On the other hand, the CNN/DailyMail dataset is created by researchers from University of Oxford to train document summarizer models. The dataset comprises 300k unique news articles sourced from the Daily Mail website along with bullet point summaries crafted by human editors. The articles encompass a wide range of subjects, encompassing politics, sports, entertainment, and beyond. Before analysis, the dataset is subjected to preprocessing to ensure the text is clean and ready for evaluation. To achieve the research objective, multiple text summarization techniques such as extractive, abstractive, and topic-based summarization are employed. The effectiveness of each technique is evaluated using ROUGE-1, ROUGE-2 and ROUGE-L scores which are widely used in extractive summarization tasks (Liu, 2019; Zhong et al., 2020; Verma, Gupta, Anil, and Chauhan, 2022) due to their focuses on content overlap, particularly emphasizing recall oriented evaluation that measures how well generated summaries align with reference summaries. On the other hand, BLEU and METEOR scores are not used for evaluation in this study because these metrics heavily used on the text generation tasks such as machine translation (Bansal, Kamper, Livescu, Lopez, and Goldwater, 2018). The findings of this research are presented in detail in the experimental results section, and the comparative analysis pipeline is visually represented in Figure 6. Text preprocessing is an essential step in preparing data for text summarization. It involves several techniques that help to clean and transform the raw text into a format that is suitable for analysis. Some of the most common preprocessing techniques include removing punctuation, converting all text to lowercase, eliminating stopwords, and performing lemmatization. Removing punctuation

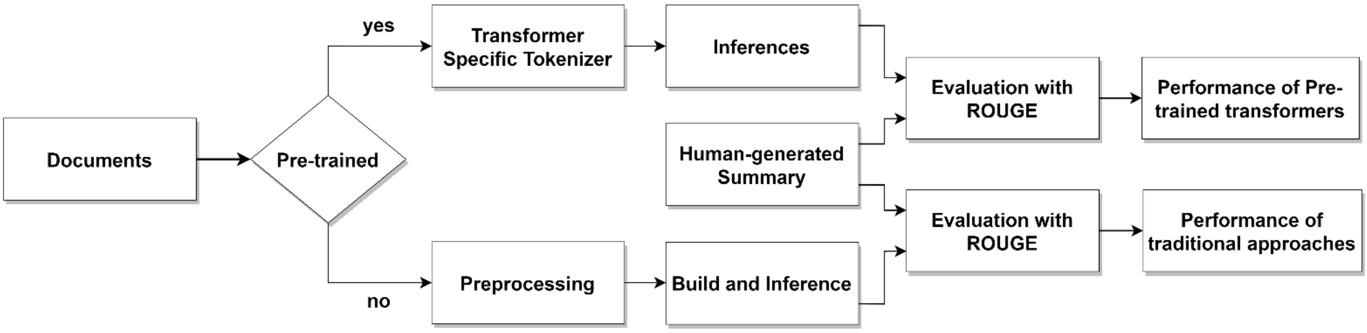

Figure 6. Text Summarization Comparative Analysis Methodology

helps to simplify the text and prevent the summarization algorithm from being confused by extraneous characters. Lowercasing the text ensures that words with different capitalizations are treated as the same, reducing redundancy in the summary. Eliminating stopwords, which are commonly used words such as "the" and "and," also helps to reduce redundancy and improve the quality of the summary. Lemmatization involves reducing words to their base form, which helps to further simplify the text and make it easier to understand. Finally, counting word frequencies helps to identify the most important and frequently used words in the document, which can inform the summarization process and ensure that the summary captures the most salient points of the original text.

In the subsequent section, the presentation of experimental findings will be encompassed, elucidating the efficacy of individual algorithms within diverse domains such as business, entertainment and sport.

5. Experimental Results

Hardware plays a crucial role in effectively training and deploying neural networks. The choice of hardware is critical as it significantly impacts the scalability and performance of deep learning models. The processing power and memory capacity of the hardware directly influence the training speed, result accuracy, and ability to handle large datasets. In this investigation, we employed a specific experimental setup comprising a GIGABYTE GeForce GTX1070 graphics processing unit, 16GB of 3000Mhz DDR4 Dual Kit random Access memory, and an INTEL Core i5 8400 2.8GHz 9MB cache 6-core central processing unit.

This study aims to evaluate the effectiveness of various text summarization techniques, namely frequency-based text summarization, latent semantic analysis-based text summarization, and graph-based text ranking. Additionally, three state-of-the-art transformer models, namely BART Large Transformer, PEGASUS Transformer, and DistilBART Transformer, are employed to benchmark the text summarization task. The benchmark evaluation is conducted on diverse domains, including business, entertainment, and sport news articles sourced from BBC. The experiments are also conducted on CNN/DailyMail news dataset which includes more technical and complex documents than BBC news dataset. By exploring these techniques and transformer models across multiple domains, this study seeks to provide valuable insights into the performance and applicability of different approaches in text summarization. The benchmark results for the business category obtained from BBC news are presented in Table 1.

Table 1. Benchmark Results for Business Category

Algorithm	ROUGE-1	ROUGE-2	ROUGE-l
Frequency Based	54.92	44.46	54.36
LSA	**55.08**	**50.13**	**59.92**
TextRank	33.41	18.83	33.08
BART large	42.23	27.43	41.04
DistilBART	44.95	29.85	43.73
PEGASUS	38.37	25.53	37.46

Based on the results achieved in Table 1, when comparing the performance of algorithms in the business category using ROUGE-1, ROUGE-2, and ROUGE-l metrics, LSA demonstrated the highest scores, indicating its strong ability to capture semantic meaning and generate summaries that align well with reference summaries. While Frequency-Based, BART Large, DistilBART, and Pegasus also performed reasonably well, outperforming TextRank, they fell short of the performance achieved by LSA. TextRank yielded the lowest scores, suggesting its limited effectiveness in generating summaries for the business category. It's worth emphasizing that the transformer models used in this context were not fine-tuned specifically for this dataset. However, they can be fine-tuned to improve performance within the specific domain. On the contrary, LSA, which delivered the highest performance in this task, is not inherently trainable. Additionally, it is important to consider the specific dataset, reference summaries, and evaluation metrics employed when evaluating algorithm performance, as these factors heavily influence the results.

Table 2 shows the benchmark outcomes from BBC news specifically for the technology category.

Table 2. Benchmark Results for Technology Category

Algorithm	ROUGE-1	ROUGE-2	ROUGE-l
Frequency Based	52.24	42.41	54.47
LSA	**54.24**	**48.07**	**58.92**
TextRank	33.20	18.63	33.55
BART large	42.72	26.83	41.43
DistilBART	45.24	28.86	44.12
PEGASUS	32.94	24.47	37.83

Considering the results achieved in Table 2, LSA once again showcased the most impressive scores, closely matching the reference summaries. Additionally, BART Large and DistilBART also produced satisfactory results when considering the ROUGE metric. However, PEGAUS and TextRank yielded unsatisfactory outcomes.

Table 3 presents the benchmark results obtained from BBC news, focusing specifically on the sport category.

Table 3. Benchmark Results for Sport Category

Algorithm	ROUGE-1	ROUGE-2	ROUGE-l
Frequency Based	49.17	41.24	51.64
LSA	**51.86**	**46.73**	**56.23**
TextRank	27.46	16.19	28.66
BART large	39.46	27.32	41.57
DistilBART	40.93	27.76	42.37
PEGASUS	37.37	29.25	40.81

Based on the outcomes presented in Table 3, LSA demonstrated exceptional performance with scores that closely resembled the reference summaries. The ROUGE metric indicated that BART Large and DistilBART also performed well, yielding satisfactory results. However, PEGAUS and TextRank, on the other hand, produced unsatisfactory outcomes.

The experiments were performed using the CNN/DailyMail dataset, which comprises a collection of documents that are more intricate and technically advanced compared to those present in the BBC News dataset. The outcomes of this particular experiment are presented in Table 4.

Table 4. Benchmark Results for CNN/DailyMail Dataset

Algorithm	ROUGE-1	ROUGE-2	ROUGE-l
Frequency Based	21.16	7.90	23.46
LSA	23.45	9.33	25.89
TextRank	16.91	4.20	17.36
BART large	**44.18**	**22.36**	**41.18**
DistilBART	43.36	21.16	40.89
PEGASUS	43.28	22.13	40.36

Based on the findings presented in Table 4, it is evident that transformer-based architectures exhibit a significantly higher level of performance compared to traditional approaches. Notably, the BART large model emerges as the frontrunner, demonstrating the most superior performance among the evaluated models. It is crucial to emphasize that these transformer-based models have undergone a process of fine-tuning on the CNN/DailyMail dataset, enabling them to better capture the relationships between sentences, particularly when compared to the traditional approach employed on the comparatively more complex dataset than BBC news. These findings illustrate that fine-tuning transformer-based architectures on a specific dataset results in significantly improved performance compared to conventional approaches.

5. Conclusion

Text summarization has become an invaluable tool for retrieving and managing information across various domains. This paper provides a comprehensive overview of text summarization techniques, covering both extractive and abstractive approaches, as well as different modeling methods such as unsupervised, supervised, and deep learning. Comparative analysis are also conducted on BBC News and CNN/DailyMail datasets. By comparing NLP transformers with traditional methods using these datasets, the study highlights their respective strengths and limitations. Surprisingly, the experimental results indicate that even the seemingly outdated LSA-based text summarization technique can yield satisfactory results in specific use cases and domains. These findings underscore the importance of considering the specific requirements and characteristics of a given domain when selecting an appropriate text summarization method. However, transformer-based architectures possess the ability to be trained and acquire a profound understanding of intricate relationships. Therefore, when trained on a specific problem, these models surpass traditional approaches in terms of performance. As the field of text summarization continues to evolve, further research and exploration are essential to enhance the effectiveness and applicability of these techniques in addressing real-world information challenges.

Conflict of Interest

We declared that there is no conflict of interest between the authors and their respective institutions. This study adheres to the ethical standards and principles governing scientific research and publishing.

Contributions of the Authors

In this study, all of the work is conducted by Yazar1.

References

Abdel-Salam, S., & Rafea, A. (2022). Performance study on extractive text summarization using BERT models. *Information*, 13(2), 67. https://doi.org/10.3390/info13020067.

Abdelaleem, N. M., Kader, H. A., & Salem, R. (2019). A brief survey on text summarization techniques. *IJ of Electronics and Information Engineering*, 10(2), 103-116.

Altmami, N. I., & Menai, M. E. B. (2022). Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*, 34(4), 1011-1028.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2018). Low-resource speech-to-text translation. *arXiv preprint arXiv:1803.09164*. https://doi.org/10.48550/arXiv.1803.0916.

Bhandari, M., Gour, P., Ashfaq, A., Liu, P., & Neubig, G. (2020). Re-evaluating evaluation in text summarization. *arXiv preprint* arXiv:2010.07100. https://doi.org/10.48550/arXiv.2010.07100.

Cagliero, L., Garza, P., & Baralis, E. (2019). ELSA: A multilingual document summarization algorithm based on frequent itemsets and latent semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 37(2), 1-33. https://doi.org/10.1145/3298987.

Cai, T., Shen, M., Peng, H., Jiang, L., & Dai, Q. (2019). Improving transformer with sequential context representations for abstractive text summarization. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I* (pp. 512-524). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32233-5_40.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165, 113679. https://doi.org/10.1016/j.eswa.2020.113679.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Advances in Information Retrieval: *27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27* (pp. 345-359). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25.

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning* (pp. 377-384). https://doi.org/10.1145/1143844.1143892.

Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121, 49-65. https://doi.org/10.1016/j.eswa.2018.12.011.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems, 28*.

Isikdemir, Y. E., & Yavuz, H. S. (2022). The scalable fuzzy inference-based ensemble method for sentiment analysis. *Computational Intelligence and Neuroscience*, 2022.

Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, 129, 200-215. https://doi.org/10.1016/j.eswa.2019.03.045.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint* arXiv:1910.13461. https://doi.org/10.48550/arXiv.1910.13461.

Li, C., Xu, W., Li, S., & Gao, S. (2018). Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 55-60). https://doi.org/10.18653/v1/N18-2009.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches* out (pp. 74-81).

Liu, S. H., Chen, K. Y., & Chen, B. (2020). Enhanced language modeling with proximity and sentence relatedness information for extractive broadcast news summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3), 1-19. https://doi.org/10.1145/3377407.

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint* arXiv:1908.08345. https://doi.org/10.48550/arXiv.1908.08345.

Liu, Y. (2019). Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*. https://doi.org/10.48550/arXiv.1903.10318.

Manojkumar, V. K., Mathi, S., & Gao, X. Z. (2023). An Experimental Investigation on Unsupervised Text Summarization for Customer Reviews. *Procedia Computer Science*, 218, 1692-1701. https://doi.org/10.1016/j.procs.2023.01.147.

Nazari, N., & Mahdavi, M. A. (2019). A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1), 121-135. https://doi.org/10.22044/jadm.2018.6139.1726.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Ramesh, A., Srinivasa, K. G., & Pramod, N. (2014). SentenceRank—a graph based approach to summarize text. In The *Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* (pp. 177-182). IEEE. https://doi.org/10.1109/ICADIWT.2014.6814680.

Ramina, M., Darnay, N., Ludbe, C., & Dhruv, A. (2020). Topic level summary generation using BERT induced Abstractive Summarization Model. In 2020 *4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 747-752). IEEE. https://doi.org/10.1109/ICICCS48265.2020.9120997.

Rodríguez-Vidal, J., Carrillo-de-Albornoz, J., Amigó, E., Plaza, L., Gonzalo, J., & Verdejo, F. (2020). Automatic generation of entity-oriented summaries for reputation management. *Journal of Ambient Intelligence and Humanized Computing*, 11, 1577-1591. https://doi.org/10.1007/s12652-019-01255-9.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint* arXiv:1704.04368. https://doi.org/10.48550/arXiv.1704.04368.

Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2(1), 1-37. https://doi.org/10.1145/3419106.

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9, 13248-13265. https://doi.org/10.1109/ACCESS.2021.3052783.

Verma, P., & Om, H. (2019). A novel approach for text summarization using optimal combination of sentence scoring methods. *Sādhanā*, 44, 1-15. https://doi.org/10.1007/s12046-019-1082-4.

Verma, S., Gupta, N., Anil, B. C., & Chauhan, R. (2022). A Novel Framework for Ancient Text Translation Using Artificial Intelligence. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 11(4)*, 411-425. https://doi.org/10.14201/adcaij.28380.

Vhatkar, A., Bhattacharyya, P., & Arya, K. (2020). Knowledge graph and deep neural network for extractive text summarization by utilizing triples. *In Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarisation* (pp. 130-136).

Wang, Q., Liu, P., Zhu, Z., Yin, H., Zhang, Q., & Zhang, L. (2019). A text abstraction summary model based on BERT word embedding and reinforcement learning. *Applied Sciences*, 9(21), 4701. https://doi.org/10.3390/app9214701.

Widjanarko, A., Kusumaningrum, R., & Surarso, B. (2018). Multi document summarization for the Indonesian language based on latent dirichlet allocation and significance sentence. In *2018 International Conference on Information and Communications Technology (ICOIACT)* (pp. 520-524). IEEE. https://doi.org/10.1109/ICOIACT.2018.8350668.

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer* *and Information Sciences*, 34(4), 1029-1046. https://doi.org/10.1016/j.jksuci.2020.05.006.

Wu, Z., Lei, L., Li, G., Huang, H., Zheng, C., Chen, E., & Xu, G. (2017). A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, 84, 12-23. https://doi.org/10.1016/j.eswa.2017.04.054.

Yao, K., Zhang, L., Luo, T., & Wu, Y. (2018). Deep reinforcement learning for extractive document summarization. *Neurocomputing*, 284, 52-62. https://doi.org/10.1016/j.neucom.2018.01.020.

Zhang, X., Wei, F., & Zhou, M. (2019). HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint* arXiv:1905.06566. https://doi.org/10.48550/arXiv.1905.06566.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.

Zheng, J., Zhao, Z., Song, Z., Yang, M., Xiao, J., & Yan, X. (2020). Abstractive meeting summarization by hierarchical adaptive segmental network learning with multiple revising steps. *Neurocomputing*, 378, 179-188. https://doi.org/10.1016/j.neucom.2019.10.019.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*. https://doi.org/10.48550/arXiv.2004.08795.