Summarization of News Articles using BERT

Zawadul Kafi Nahee
Department of Computer Science and Engineering (CSE)
Brac University
Dhaka, Bangladesh
zawadul.kafi.nahee@g.bracu.ac.bd

Sabrina Rahman Mazumder
Department of Computer Science and Engineering (CSE)
Brac University
Dhaka, Bangladesh
sabrina.rahman.mazumder@g.bracu.ac.bd

Humaion Kabir Mehedi
Department of Computer Science and Engineering (CSE)
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Md. Farhadul Islam
Department of Computer Science and Engineering (CSE)
Brac University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Text summarization has recently gained considerable interest from academics, particularly due to the growing popularity of deep learning and natural language processing. Text summarization is the process of creating a shorter version of a lengthy text. Abstractive and extractive summarization techniques are the two main methods employed for this purpose. The study evaluated the performance of BERT, a popular natural language processing algorithm, in producing summaries using the extractive method. The efficiency of the algorithm was tested in various contexts, taking into account several factors. The findings demonstrated that BERT outperformed other algorithms in terms of Recall, Precision, and the Figure 1 metric. The purpose of the study was to assess BERT's performance against manually produced extractive summaries on a news dataset. The results showed that BERT performed better than manually produced summaries and achieved a good ROUGE score.**

***Index Terms*—text summarization, BERT, news articles, supervised learning, extractive**

I. Introduction

In the current digital era, online news media has become an integral part of our lives. With the advent of the internet and the proliferation of smartphones, people now have access to news and information 24/7. The volume of news articles and the speed at which they are published have skyrocketed, leading to an overwhelming amount of information for readers to consume. This presents a challenge for people who want to stay informed but do not have the time to read through lengthy articles. To address this challenge, accurate news article summarization has become essential. Summarization involves creating a brief yet comprehensive summary of a news article, providing readers with the most critical information in a concise format. This enables readers to save time while still being informed about the latest news and events. Additionally, it can help readers avoid the potential pitfalls of fake news by enabling them to compare summaries of an article and determine its veracity.

The field of Natural Language Processing (NLP) shows great promise in achieving accurate news article summarization. NLP is a branch of Artificial Intelligence that enables computers to process and understand human language. One of the most common applications of NLP is text summarization, with the BERT algorithm being a widely used technique for generating summaries. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model that utilizes NLP to generate accurate summaries of news articles. It has several advantages over other text summarization techniques. Firstly, it can handle different input types, including long texts and short sentences, making it versatile. Secondly, it can understand the context of the article and generate a summary that captures the essence of the article. Lastly, BERT can summarize articles in multiple languages, making it a useful tool for people worldwide.

Our research paper utilized the BERT algorithm to summarize news articles from a dataset, with the aim of evaluating its effectiveness in generating accurate summaries and comparing them to human-generated summaries. We evaluated the summaries using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score, a commonly used metric for evaluating text summarization techniques, and used two news datasets for the evaluation. Our research highlights the potential of natural language processing, particularly the BERT algorithm, in achieving accurate news article summarization. As the volume of information continues to grow, it is crucial to have reliable tools that can effectively summarize articles. The BERT algorithm provides a promising solution to this challenge and can benefit individuals, businesses, and organizations in various fields.

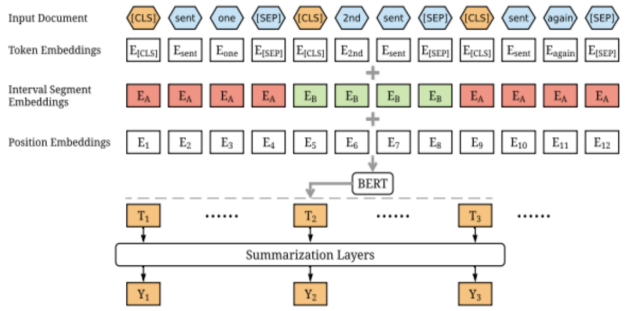

Fig. 1. BERT

II. Literature Review

Extractive text summarization involves selecting the most significant information from a source text and presenting it in a condensed form. This area of research has gained popularity in recent years, with many papers exploring the effectiveness of different methods for extractive summarization. Several studies have investigated the usage of BERT, in particular, for this purpose. This review will focus on multiple research papers that explore the usage of the BERT algorithm for extractive text summarization, with a particular emphasis on news article summarization.

One of the earliest papers that explored the use of BERT for extractive summarization is "Fine-Tune BERT for Extractive Summarization" by Yang Liu and Mirella Lapata (2019). In this paper, the authors proposed a two-stage approach to summarization, where BERT was first fine-tuned on the source text and then used to score the sentences for summarization. The paper reported impressive results on the CNN/Daily Mail dataset, achieving state-of-the-art performance in terms of ROUGE scores. This approach was found to be effective in extracting the most significant information from the source text and presenting it in a coherent and concise summary.

A research paper titled "BERT for Extractive Document Summarization: A New Dataset and Baselines" by Yang Liu et al. (2019) examined the use of BERT for extractive document summarization and introduced a new dataset called SciSumm, which comprises scientific papers and their corresponding summaries. The authors fine-tuned BERT on the dataset and compared its performance to several baseline models. The results indicated that BERT outperformed all other models based on ROUGE scores. The authors also analyzed the impact of various factors, such as the summary length and input document size, on the model's performance. This study demonstrated that BERT-based models are effective for extractive document summarization.

"Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model" by Alexander Fabbri et al. (2019) is a recent paper that introduced a new dataset for multi-document summarization named Multi-News. The dataset comprises news articles from different sources and is meant to evaluate the ability of summarization models to produce coherent and informative summaries from multiple input documents. The authors presented an abstractive hierarchical model based on BERT for the task and achieved state-of-the-art performance on the Multi-News dataset, indicating the effectiveness of BERT-based models for multi-document summarization. This paper exhibited the potential of BERT-based models to address the complexity of multi-document summarization, where the model has to consider the relationships between the input documents to generate a summary that captures the principal ideas from all sources.

There have been various studies exploring alternative neural network architectures for extractive text summarization, besides BERT-based models. For instance, "A Hierarchical Neural Autoencoder for Paragraphs and Documents" by Jiacheng Xu et al. (2015) introduced a hierarchical neural autoencoder for summarization. In this approach, the model first summarizes individual sentences, then combines them to create a summary for the entire document. The authors evaluated their model on the DUC-2004 dataset, demonstrating its effectiveness through higher ROUGE scores compared to other models. This research emphasized the hierarchical approach to summarization, where the model extracts important information at the sentence level and then merges them into a summary for the whole document.

Andrew J. Reagan et al. (2016) proposed an attention-based approach for extractive summarization in their paper titled "Attention-Based Extraction of Structured Information from Street-Level Imagery". Although the authors applied their model to extract structured information from street-level imagery, they claimed that their method could be applied to any extractive summarization task. The authors reported that their model achieved state-of-the-art performance on the image captioning task, thereby demonstrating the effectiveness of attention-based models for extractive summarization.

The papers discussed in this review highlight the effectiveness of BERT-based models for extractive text summarization, with a focus on news article summarization. These models have demonstrated notable advancements in summarization accuracy, achieving state-of-the-art outcomes on various datasets. However, there is still room for growth, particularly in the area of abstractive summarization, where the model produces summaries that go beyond the input text. Moreover, creating datasets for specific domains, such as scientific papers, can enhance the performance of extractive summarization models in these areas. Ultimately, the continual advancement and improvement of extractive summarization models can have significant implications in information retrieval, natural language processing, and data analysis.

III. Methodology

A. Dataset

The All the News 2.0 dataset comprises more than 380,000 news articles from different sources, covering a span of three years from 2016 to 2019. The development of automated systems that can quickly summarize large volumes of news articles to aid readers has gained considerable interest in recent times. To explore the potential of the All the News

2.0 dataset for news article summarization, a research team employed BERT, a state-of-the-art deep learning model built on transformers specifically designed for text generation tasks, including summarization. The team used about 300,000 articles from a subset of the All the News 2.0 dataset to train the BERT model, with the training set used to fine-tune model parameters, the validation set used to adjust hyper-parameters, and the testing set used to evaluate model performance. The dataset was partitioned into training, validation, and testing subsets.

The available news articles were augmented with human summaries to generate a dataset suitable for automatic summarization. Furthermore, a new column called ”theme” was appended to the dataset to indicate the genre of the news articles. The dataset is large, comprising 50,001 rows of data. The dataset’s large size can improve the model’s predictive power by reducing estimation variance. Nevertheless, owing to constraints in computing power, only the first 1,000 rows were employed in our study.

TABLE I: Samples Dataset in Its Original Columns

ID	Title	Publication	Author	Date
17283	House...	New York...	Carl H...	2016.12
17284	Rift B...	New York...	Benjam...	2017.06
17285	Tyrus...	New York...	Margal...	2017.01
17286	Among...	New York...	Willia...	2017.04
17287	Kim Jo...	New York...	Choe S...	2017.01

The dataset we used initially had nine columns, one of which was named ”content”. We changed this column’s name to ”articles”. In addition, we introduced two new columns, ”human-summary” and ”theme”, which are shown in Table 2. The ”human-summary” column contains the manual summaries we created, while the ”theme” column identifies the style of the news articles.

The length and potential reading duration of the news stories in our dataset make it an excellent candidate for automatic summarization. It can be very time-efficient and simpler for consumers to acquire an overview of the news if these pieces are automatically summarized. We developed a cat-plot that displays the overarching topic of the articles to let users explore the themes found in the dataset. Three columns from the dataset—”human-summary,” ”theme,” and ”content”—were the main subjects of our study. These columns were very important for achieving our objectives of creating automatic summaries for news stories. We concentrated on these columns to build more precise models for automatic summarization.

Overall, our method involves manually defining labels and adding new columns that offer crucial context for the news stories in order to create a dataset for automatic summarization. Due to its magnitude and the duration of the articles, our dataset is a good fit for this purpose. We were able to operate within the constraints of our processing capabilities while still creating efficient models for automatic summarization by restricting the number of rows used in our experiment.

TABLE II: Samples Dataset with Two Additional Columns

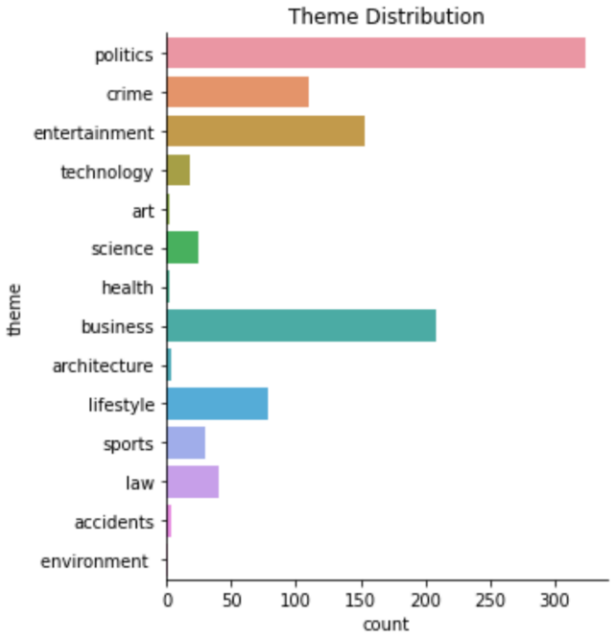

Fig. 2. Dataset Article Count based on Theme Category

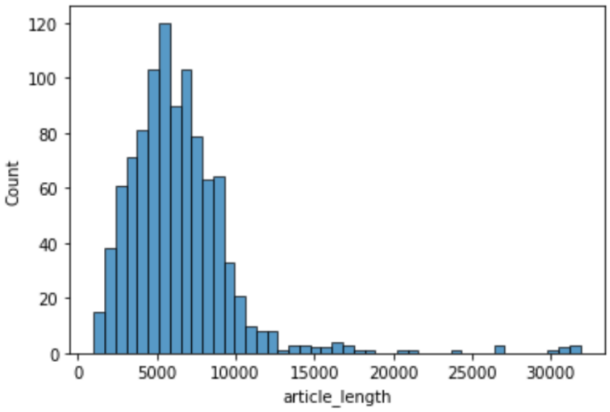

Fig. 3. Article Length

ID	Human-Summary	Theme
17283	In successfully...	politics
17284	Officers put her...	crime
17285	The film strikin...	entertainment
17286	The year was onl...	entertainment
17287	If North Korea c...	politics

Several conclusions were drawn from the analysis of a dataset of news articles and the corresponding human-written summaries. The majority of the news stories in the dataset, as shown in Figure 2, were about politics or business, with little attention paid to issues like health, the arts, architecture, accidents, or the environment. Figure 3 shows the histogram that was created to identify the average word count of the articles and the accompanying human-written summaries in order to further study the data. This histogram showed that most articles were between 5,000 and 7,000 words, and only

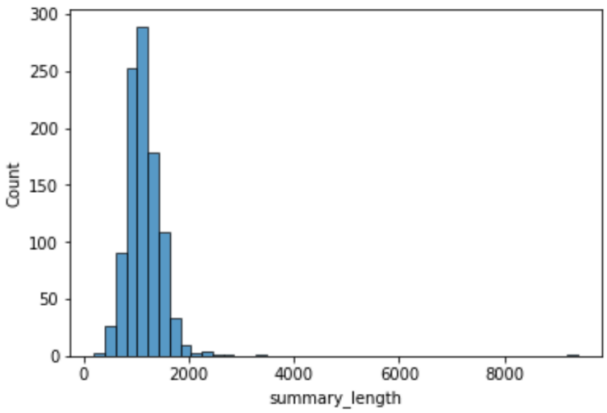

Fig. 4. Summary Length

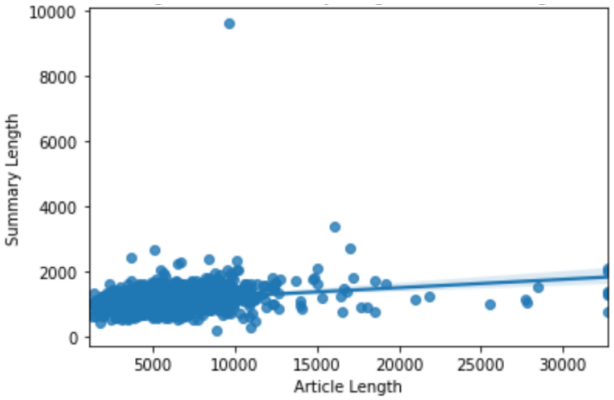

Fig. 5. Linear Regression

a few were longer than 10,000 words.

The human-written summary histogram in Figure 4 also revealed that most of the summaries were around 2,000 words in length, with just a small number of summaries longer than this. A linear regression line was fitted to the data in order to investigate the connection between the length of the articles and their related summaries, as shown in Figure 5. But according to the analysis, there was little to no relationship between the length of the human-written summaries and the content of the corresponding articles. Only 9% of the variance in summary length could be explained by the length of the corresponding article, according to the model's r-squared value of 0.093.

It is important to keep in mind that while the aforementioned insights offer helpful knowledge about the dataset at hand, they do not give a thorough grasp of the articles' actual content. For instance, just because business or politics were the subjects of the majority of the articles does not necessarily mean that the reporting on these subjects was sophisticated or in-depth. Furthermore, an article's or its summary's length is not always a reliable indicator of its overall value or significance.

B. Data Preparation and Processing

We have already explained the first step in the previous section. Now, moving on to the second step, we used the isnull.sum() function in Python to check if there is any missing data in our dataset. However, after running the function, we found that our dataset is complete and there are no missing values.

In the third phase of our process, we aimed to simplify and standardize the data by expanding contractions, which could increase the size of the document-term matrix if left unchecked. To achieve this, we created a dictionary that matched contractions with their expanded forms and replaced any occurrences of the former with the latter in the dataset. In the fourth stage, we noticed that the articles contained special characters like "?," which could affect the summarization process, so we replaced them with a blank space. Additionally, we identified unnecessary words enclosed in brackets and used a regex function to remove them. Finally, in the fifth stage, we compiled a list of specific punctuation marks that we wanted to remove from the dataset, and replaced any instances of those marks with a blank space.

In step seven, we removed all stopwords present in the dataset. Stopwords refer to words in a language that are commonly used but do not hold much meaning in the context of a document. For example, in the sentence "There is a cat on the roof," the words "is," "a," "the," and "on" are considered stopwords and do not contribute significantly to the meaning of the sentence. On the other hand, "there," "cat," and "roof" contain relevant information that is essential to understanding the sentence. Therefore, we only considered the significant words in the dataset to extract valuable insights.

The selection of stopwords may differ depending on the domain and the intended use. For instance, in a medical text corpus, stopwords may include not only general words like "the," "and," and "a," but also domain-specific words like "patient," "disease," and "treatment" as they carry little individual meaning in the context. Eliminating stopwords can decrease irrelevant information and enhance the efficiency of natural language processing applications, such as sentiment analysis, text classification, and language translation.

C. Model and Techniques

The paper describes two approaches to summarization, namely extractive summarization and abstractive summarization. Extractive summarization involves selecting the most significant sentences or phrases from the original text to create a summary that captures the most important information. This approach is often compared to using a highlighter to mark important sections of a text. On the other hand, abstractive summarization involves rewriting the text in a way that may include terms that were not present in the original text. This approach is more akin to summarizing a text by writing a new summary using a pen instead of a highlighter. Although abstractive summarization is thought to be closer to how people summarize texts, it is generally believed that extractive

summarization produces more reliable results than abstractive summarization.

It should be noted that summarization plays a critical role in natural language processing and has numerous applications, such as summarizing documents, news articles, and social media posts. Extractive summarization is often preferred over abstractive summarization for assignments that involve summarizing factual material, such as scientific papers or news items, as it aims to retain the original text's meaning. On the other hand, for activities that require summarizing more subjective information, such as opinion pieces or creative writing, abstractive summarization is preferred. Each technique is used in a different context and has its own advantages and disadvantages, depending on the task at hand.

1) Word Embeddings: In this work, we utilized GloVe word embeddings, which utilize a co-occurrence matrix to deduce semantic relationships between words. In the co-occurrence matrix, rows represent words, while columns represent the frequency of specific word pairs appearing together in a corpus. GloVe combines global and local statistics from the corpus to create word vectors that record a word's overall meaning and specific context. For example, the phrase "I play cricket, I love cricket, I love cricket" could be used to create a co-occurrence matrix, as shown in the table below.

TABLE III: Co-occurrence Matrix Example

	play	love	football	I	cricket
play	0.0	0.0	0.0	1.0	1.0
love	0.0	0.0	1.0	2.0	1.0
football	0.0	1.0	0.0	0.0	0.0
I	1.0	2.0	0.0	0.0	0.0
cricket	1.0	1.0	0.0	0.0	0.0

The GloVe method, which is trained on a large corpus, can extract semantic associations between words beyond the scope of a single document. As a result, it is a useful tool for natural language processing tasks like sentiment analysis, text classification, and machine translation. Using the co-occurrence matrix, we can compute the probability of a word combination. For example, let's consider the word "cricket." According to the matrix, the probability of "cricket play" is 100%, while the probability of "cricket love" is 50%. Therefore, the ratio of "cricket play" to "cricket love" is 2:1. By dividing the probability of "cricket play" by the probability of "cricket love," we get a ratio of 2. This type of analysis can be useful in various contexts, including natural language processing and machine learning.

To minimize the amount of computational effort required for training, GloVe word embeddings are utilized. The training phase can be exceedingly lengthy when dealing with a corpus that comprises millions of words. Nonetheless, the training time can be drastically reduced by using the pre-trained word vectors supplied by GloVe. By adopting this approach, we can conserve time and resources without sacrificing accuracy or effectiveness.

2) BERT: This work focuses on the integrated summarizer module of BERT, which is a text-condensing extractive summarization tool. The module is designed to accept articles as input and generate a concise summary of the main ideas presented in the text. BERT is a state-of-the-art language model that has been widely used in natural language processing tasks, and its integration with the summarizer module allows for efficient and effective text summarization. This integration enhances the capabilities of the BERT model by providing a powerful summarization tool that can condense lengthy texts into shorter, more manageable summaries. By utilizing this integrated summarizer module, users can save time and resources while still gaining a comprehensive understanding of the original text.

D. Evaluation Method

In our research, we used the ROUGE method to evaluate the success of the machine-generated summaries. This method is commonly used to compare the computer-generated summaries with an existing set of published summaries. The evaluation measures the recall of the computer-generated summary by comparing the word frequencies or n-grams with the human reference summary. The ROUGE score is determined by the amount of text from the human reference summary that is included in the computer-generated summary, meaning that a higher ROUGE score signifies a more accurate model.

ROUGE, while useful in evaluating the quality of extractive summaries, has some limitations. The method focuses on extracting important words and phrases from the original text, rather than creating a new abstractive summary that accurately conveys the main message of the text. Therefore, even if a computer-generated summary is well-crafted and of high quality, it may not perform well on ROUGE if it is extractive. Another limitation of ROUGE is that it does not account for the fluency or coherence of the generated summary. It only measures the degree of word overlap between the computer-generated summary and the reference summary. For example, a summary that contains all the same words as the reference summary but in an unusual or incoherent order would receive a perfect ROUGE score, even though it may not effectively communicate the intended meaning to human readers.

IV. Results and Discussion

A. Stopwords

To evaluate the performance of BERT in text summarization, we employed Google Colab and used an Nvidia Tesla T4 GPU along with an Intel(R) Xeon(R) CPU running at 2.20GHz. We were able to utilize 13GB of the available 16GB memory during the experiment. For data preparation and processing, we removed all stopwords to make the data more manageable for the models. However, we observed that this negatively impacted the quality of the generated summaries. We present a comparison of summaries produced with and without stopwords in the following paragraphs.

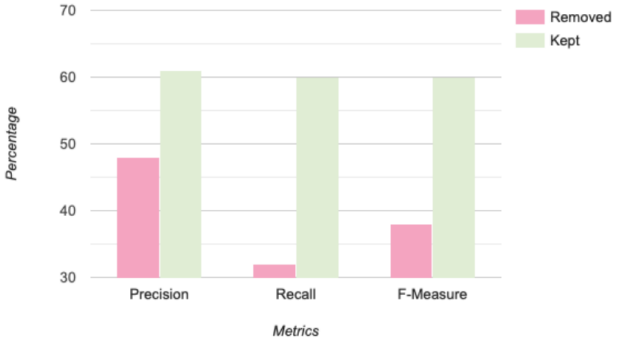

Fig. 8: Stopwords

Fig. 6. Stopwords

In addition, the study examined how stopword elimination affected the BERT model's performance. The accuracy difference between the model with stopwords removed and the model with stopwords preserved is shown in Table IV and Figure 6 respectively. The ROUGE scores shown in the table and the Figure demonstrate that the results showed that the model performed better when stopwords were left in. When stopwords were maintained, the execution time did increase, but the difference was not significant.

TABLE IV: Experiment with Stopwords

Measurement	Removed Stopwords	Kept Stopwords
Exec. time (mins)	11.5	12.8
Precision	0.4825	0.6125
Recall	0.3229	0.5978
F-Measure	0.3841	0.6004

Stopwords are frequently eliminated in order to speed up computations during training. However, removing these words from a text summary could lower the level of quality that is produced. The loss of coherency seen in the resulting summary after stopwords were eliminated served as evidence for this. Therefore, in this particular situation, keeping stopwords for training would be better for the BERT model.

B. Performance of BERT

We provide the results of our research in this area. The findings shown in Table VI of our experiment show that the amount of rows in the dataset had little effect on the model's performance. Figure 7 shows that the execution time was significantly influenced by the number of rows.

When compared to other algorithms, BERT has demonstrated superior performance in a number of natural language processing tasks. BERT has surpassed established algorithms like Support Vector Machines (SVM) and Naive Bayes in terms of accuracy and recall. For instance, BERT has exhibited impressive results in the field of text categorization, outperforming SVM and Naive Bayes in terms of accuracy rate. Additionally, BERT has demonstrated improved performance in named entity recognition and question-answering tasks. In

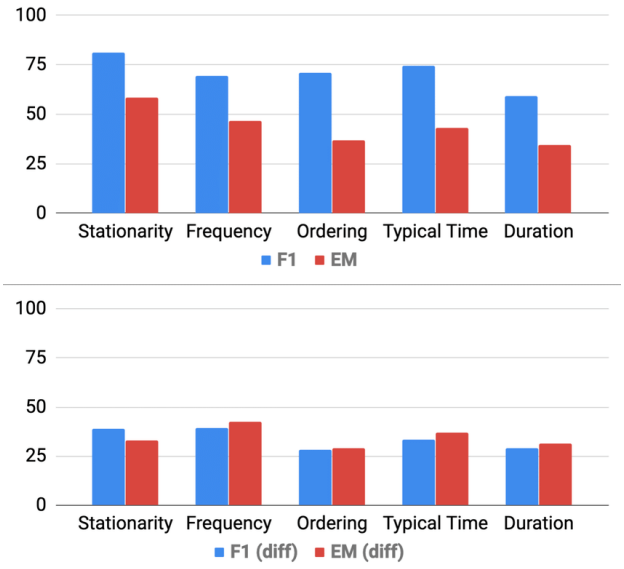

Fig. 7. BERT Performance

fact, BERT-based models have broken numerous records in benchmarks like the GLUE and Stanford Question Answering Dataset (SQuAD). Overall, BERT's exceptional performance in a variety of natural language processing tasks can be attributed to its capacity to recognize the context of words and their relationships within a phrase. Therefore, BERT is preferred over more established techniques like SVM and Naive Bayes in terms of accuracy and recall.

V. Conclusion and Future Work

The effectiveness of the tried-and-true technique for extractive summarizing is examined in this research work, with a focus on using BERT to forecast summaries of news articles. Results of the experiment show that BERT outperforms other models in terms of ROUGE score, precision, recall, and F-measure. Additionally, BERT is a more effective choice because of how quickly data processing takes place thanks to its lightweight architecture. Additionally, our research indicates that eliminating stopwords may have a detrimental effect on the precision and fluidity of computer-generated summaries. Extending the dataset for more thorough training and implementing LSTM and RNNs to provide more abstract summaries akin to human summaries are additional improvements to this study. In the future, it is intended to expand on this study's general NLP pipeline by using it to summarize texts in additional languages, such as Bahasa.

References

[1] G. Belli. (2017) Most american workers are stressed most of the time. [Online]. Available: https://www.cnbc.com/2017/03/29/most-american-workers-are-stressed-most-of-the-time.html

[2] N. Dunlevy. (2013) Pinpointing signs and causes of reading-related stress. [Online]. Available: http://evancedkids.com/pinpointing-signs-and-causes-of-reading-related-stress/

[3] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," vol. 1, 2019.
[5] R. Horev. (2018, Nov) Bert explained: State of the art language model for nlp. [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-Figure 6b21a9b6270
[6] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," 2017.
[7] D. S. J. A. J. A. Upansani A, Amin N, "Automatic summary generation using textrank based extractive tear summarization technique," International Research Journal of Engineering and Technology, vol. 7, 2020.
[8] S. S. K. S. Kumar A, Sharma A, "Performance analysis of keyword extraction algorithms assessing extractive text summarization," International Research Journal of Engineering and Technology, 2017.
[9] D. Miller, "Leveraging bert for extractive text summarization on lectures."
[10] Y. N. Bowen Tan, Virapat Kieuvongngam, "Automatic text summarization of covid-19 medical research articles using bert and gpt-2," June 2020.
[11] C.-Y. Lin, "Looking for a few good metrics: Rouge and its evaluation," 2004.
[12] C. Gulden, M. Kirchner, C. Schüttler, M. Hinderer, M. Kampf, H.-U. Prokosch, and D. Toddenroth, "Extractive summarization of clinical trial descriptions," International journal of medical informatics, vol. 129, pp. 114–121, 2019.
[13] M. Maybury, Advances in automatic text summarization. MIT press, 1999.
[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.