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*—Recent years have seen a significant increase in academic interest in text summarization, notably with the rise of deep learning and NLP. The practice of condensing a large text is known as text summarization. The two basic strategies used for this are abstractive and extractive summarization. In order to create summaries using the extractive technique, the study assessed BERT's performance, a well-known natural language processing algorithm. The algorithm's effectiveness was examined in numerous scenarios while taking a number of things into account. The results showed that BERT performed better in terms of Recall, Precision, and the Figure 1 metric than other algorithms. The goal of the study was to compare the effectiveness of BERT with manually created extractive summaries on a news dataset. The outcomes demonstrated that BERT outperformed hand crafted summaries and received a favorable ROUGE score.**

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

I. Introduction

Online news media has a significant impact on our lives in the contemporary digital era. People may now get news and information around-the-clock thanks to the internet and the widespread use of cellphones. There is an excessive quantity of information available to readers as a result of the rapid increase in the number and speed of news item publication. Those who wish to keep informed but lack the time to go through lengthy articles may find this to be a difficulty. Accurate news article summary is now crucial to overcoming this problem. In order to give readers the most important information in a clear and succinct manner, summarizing entails writing a concise yet thorough summary of a news piece. This allows readers to save time while still staying up to date on the most recent events and news. Additionally, by allowing readers to compare article summaries and assess an article's veracity, it can help readers avoid the potential pitfalls of false information.

Accurately summarizing news articles is a goal that the area of natural language processing (NLP) shows considerable promise in fulfilling. NLP is a subfield of AI that gives computers the ability to process and comprehend human language. Text summarization is one of the most frequently utilized NLP applications, and the BERT algorithm is a popular method for doing so. A pre-trained deep learning model called BERT (Bidirectional Encoder Representations from Transformers) makes use of NLP to accurately summarize news stories. Compared to other text summary methods, it provides a number of benefits. First of all, it is adaptable since it can handle a variety of input formats, including large texts and small phrases. Second, it can comprehend the article's context and provide a summary that encapsulates the substance of the piece. Last but not least, BERT is a beneficial tool for individuals all around the world because it can summarize content in several languages.

In order to assess the accuracy of the summaries it produced and compare them to summaries created by humans, our research study used the BERT algorithm to summarize news items from a dataset. Using two news datasets, we assessed the summaries using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score, a standard measure for assessing text summarizing methods. Our study shows how effective news article summaries can be produced using natural language processing, particularly the BERT algorithm. It is essential to have trustworthy tools that can efficiently summarize articles when the amount of information keeps increasing. This problem has a possible answer in the BERT algorithm, which can help people, companies, and organizations across a range of industries.

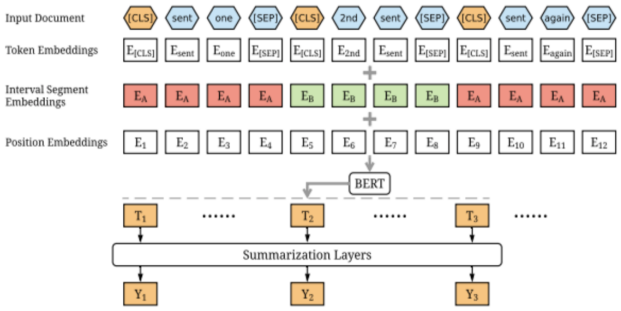

Fig. 1. BERT

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

More than 380,000 news stories from various sources make up the All the News 2.0 dataset, which spans three years from 2016 to 2019. Recently, there has been a lot of interest in the creation of automated systems that can swiftly summarize several news items for the benefit of readers. BERT is a cutting-edge deep learning model based on transformers particularly

developed for text production tasks, including summarization. A research team used it to investigate the possibilities of the All the News 2.0 dataset for news article summarizing. To train the BERT model, the researchers employed around 300,000 articles from a portion of the All the News 2.0 dataset. The training set was used to alter model parameters, the validation set to modify hyper-parameters, and the testing set to assess model performance. The dataset was divided into subgroups for training, validation, and testing.

To create a dataset appropriate for automated summarizing, the existing news items were supplemented with human summaries. In addition, a brand-new column named ”theme” was added to the dataset to represent the news items’ genre. The dataset is big, including 50,001 rows of information. Large datasets can increase the predictive power of a model by lowering estimation variance. However, due to computing power limitations, only the first 1,000 rows were used 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 news stories in our dataset are a great candidate for automated summarization due to their length and possible reading time. If these articles are automatically summarized, it may save customers a lot of time and make it easier for them to get a general sense of the news. To enable visitors to explore the themes discovered in the dataset, we created a cat-plot that displays the articles’ main topic. The three columns ”human-summary,” ”theme,” and ”content” from the dataset served as the focus of our investigation. These columns were crucial to reaching our goals of automatically summarizing news items. In order to create more accurate models for automated summarization, we focused on these columns.

In order to construct a dataset for automatic summarization, our technique entails manually defining labels and adding new columns that provide essential context for the news pieces. Our dataset is an excellent fit for this use because of its size and the length of the articles. By limiting the amount of rows utilized in our experiment, we were able to work within the limitations of our processing power while still developing effective models for automated summarization.

TABLE II: Samples Dataset with Two Additional Columns

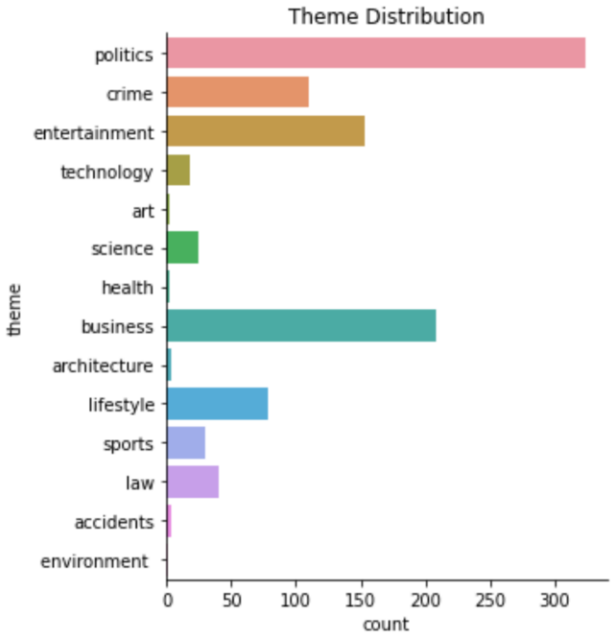

Fig. 2. Dataset Article Count based on Theme Category

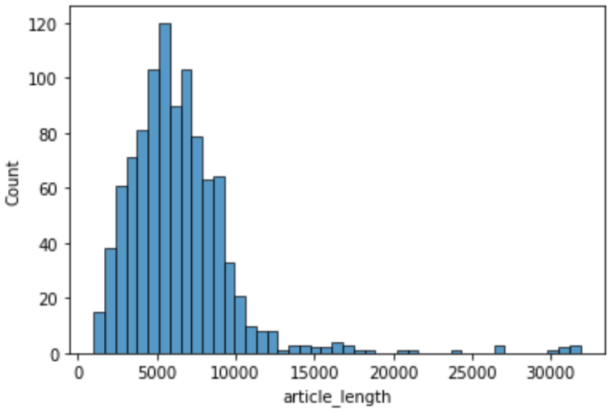

Fig. 3. Article Length

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

The study of a dataset of news stories and the related human-written summaries led to a number of results. Politics and business dominated the news items in the dataset, as seen in Figure 2, with little focus being placed on topics like health, the arts, architecture, accidents, or the environment. The average word count of the articles and the associated human-written summaries were calculated in order to further analyze the data, as shown in Figure 3. According to this histogram, only a small percentage of articles were longer than 10,000 words, with the majority falling between 5,000

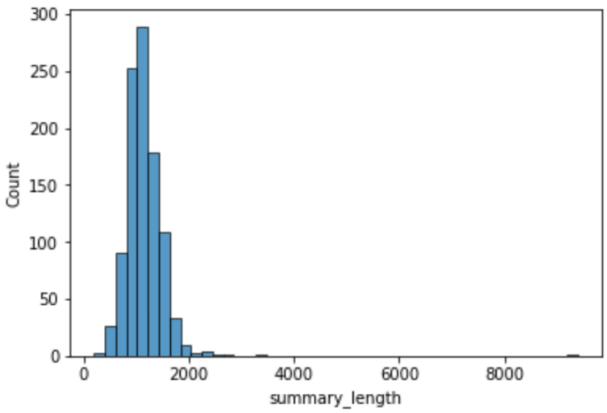

Fig. 4. Summary Length

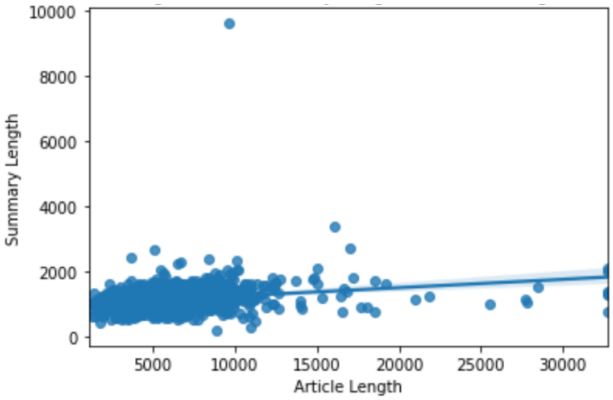

Fig. 5. Linear Regression

and 7,000 words.

The human-written summary histogram in Figure 4 likewise showed that most summaries were no longer than 2,000 words, with only a few exceeding this word count. In order to analyze the relationship between the length of the articles and their associated summaries, a linear regression line was fitted to the data, as seen in Figure 5. However, the analysis found that there was little to no correlation between the content of the corresponding articles and the length of the human-written summaries. The model's r-squared value of 0.093 indicated that only 9% of the variance in summary length could be accounted for by the length of the associated article. While the aforementioned insights provide useful knowledge about the available dataset, it is important to keep in mind that they do not provide a thorough understanding of the actual content of the articles. For example, simply because business or politics dominated the bulk of the stories does not automatically imply that the reporting on these topics was intelligent or in-depth. Additionally, the length of an article or its summary is not always a good indication of its overall worth or significance.

B. Data Preparation and Processing

The first step was already covered in the section before this one. In the second stage, we used Python's isnull.sum() method to see if any of the values in our dataset were missing. We discovered, however, that our dataset is full and contains no missing values after executing the method.

By increasing contractions, which may, if left unchecked, result in an increase in the size of the document-term matrix, we wanted to standardize and simplify the data in the third stage of our procedure. To do this, we built a vocabulary that compared contractions with their extended equivalents, and we then swapped out any instances of the former in the dataset with the latter. We discovered in the fourth step that the articles had special characters like "?," which would interfere with the summary procedure, so we substituted a blank space for them. Furthermore, we found extra words encased in brackets and eliminated them using the regex function. The fifth and final stage involved creating a list of particular punctuation marks we intended to exclude from the dataset and then replacing all occurrences of those marks with blank spaces.

We eliminated all stopwords from the dataset in step seven. Stopwords are words that are often used yet have limited significance when employed in the context of a document. In the statement "There is a cat on the roof," for instance, the words "is," "a," "the," and "on" are regarded as stopwords since they don't significantly add to the content of the sentence. On the other hand, the words "there," "cat," and "roof" are pertinent and necessary for comprehending the statement. In order to gain useful insights, we only took into account the significant terms in the dataset. Depending on the domain and intended application, several stopwords may be chosen. For instance, stopwords in a corpus of medical texts may not just include generic terms like "the," "and," and "a," but also subject-specific terms like "patient," "disease," and "treatment" since they have no intrinsic value outside of the context. Stopword elimination can reduce unnecessary data and improve the effectiveness of NLP applications including sentiment analysis, text categorization, and language translation.

C. Model and Techniques

The two methods of summarizing discussed in this work are extractive summarization and abstractive summarization. In order to construct an extractive summary that effectively summarizes the original material, the most significant sentences or phrases must be chosen. This strategy is sometimes compared to underlining key passages in a manuscript. As opposed to this, abstractive summarization entails rewriting the text using concepts that might not have been present in the original. Instead of utilizing a highlighter, this method is more equivalent to producing a fresh summary of the text. Extractive summary is regarded to give more trustworthy findings than abstractive summarization, despite the fact that abstractive summarization is seen to be closer to how individuals summarize texts.

It should be emphasized that summary has several uses, including summarizing papers, news stories, and social media postings, and that it is essential to natural language processing. For tasks that call for summarizing factual content, such scientific papers or news articles, extractive summary is frequently favored over abstractive summarization since

it seeks to preserve the original text's meaning. Conversely, abstractive summarization is preferred for tasks requiring the summarization of more subjective information, such as opinion pieces or creative writing. Depending on the work at hand, each approach is employed in a distinct context and has its own benefits and drawbacks.

1) Word Embeddings: In this study, we used GloVe word embeddings, which infer semantic associations between words using a co-occurrence matrix. In the co-occurrence matrix, columns reflect the frequency of particular word pairs appearing together in a corpus, whereas rows represent words. GloVe uses a combination of global and local data from the corpus to produce word vectors, which capture both the general meaning and particular context of a word. For instance, a co-occurrence matrix for the sentence "I play cricket, I love cricket, I love cricket" may be constructed, as seen 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 approach can extract semantic links between words outside the bounds of a single document since it is trained on a vast corpus. It is therefore a helpful tool for NLP tasks including sentiment analysis, text categorization, and machine translation. We may determine a word combination's likelihood using the co-occurrence matrix. Let's use the word "cricket" as an example. The matrix states that while "cricket play" has a 100% likelihood, "cricket love" has a 50% chance of happening. The proportion of "cricket play" to "cricket love" is therefore 2:1. The probability of "cricket play" is divided by the likelihood of "cricket love," yielding a ratio of 2. This kind of study can be helpful in a variety of settings, such as machine learning and natural language processing.

GloVe word embeddings are applied to reduce the computing effort needed for training. When working with a corpus that contains millions of words, the training step may be quite drawn out. However, by utilizing the pre-trained word vectors provided by GloVe, the training time can be significantly decreased. This strategy allows us to save time and money without compromising precision or efficiency.

2) BERT: The integrated summarizer module of BERT, a text-condensing extractive summarizing tool, is the subject of this study. The module is made to take in articles as input and produce a succinct summary of the key concepts discussed in the text. Modern language models like BERT are often utilized in jobs involving natural language processing, and their incorporation with the summarizer module enables rapid and accurate text summarizing. By offering a potent summary tool that can reduce extensive texts into shorter, more digestible summaries, this integration expands the BERT model's potential. Users may save time and money while still acquiring a thorough grasp of the original material by making use of this integrated summarizer module.

D. Evaluation Method

In order to assess the effectiveness of the computer-generated summaries, we employed the ROUGE approach in our study. This technique is frequently used to contrast computer-generated summaries with a library of previously published summaries. By contrasting the word frequencies or n-grams with the human reference summary, the assessment gauges how well the computer-generated summary is remembered. A higher ROUGE score indicates a more accurate model since it is based on how much language from the human reference summary is contained in the computer-generated summary.

ROUGE has various limitations, although being helpful in assessing the caliber of extractive summaries. Instead of writing a fresh abstractive summary that faithfully captures the core idea of the book, the technique concentrates on highlighting key words and phrases from the original text. Therefore, if a computer-generated summary is extractive, it might not perform well on ROUGE even if it is well-written and of high quality. ROUGE also has the drawback of not taking into account the coherence or fluency of the produced summary. Only the degree of word overlap between the reference summary and the computer-generated summary is measured. Even though it would not properly convey the intended meaning to human readers, a summary that uses all the same terms as the reference summary but in an odd or illogical arrangement will obtain a flawless ROUGE score.

IV. Results and Discussion

A. Stopwords

We used Google Colab and an Nvidia Tesla T4 GPU combined with an Intel(R) Xeon(R) CPU operating at 2.20GHz to assess BERT's performance in text summarization. During the experiment, we were able to use 13GB of the 16GB of memory that was available. To make the data more comprehensible for the models, we eliminated all stopwords throughout the preparation and processing of the data. However, we noticed that this had a detrimental effect on the summaries that were produced. In the paragraphs that follow, we compare summaries created both with and without stopwords.

The study also investigated the performance impact of stopword deletion on the BERT model. Table IV and Figure 6 respectively display the accuracy disparity between the model with stopwords deleted and the model with stopwords kept. The findings indicated that the model performed better when stopwords were included, as seen by the ROUGE scores in the table and Figure. The execution time did rise when stopwords were kept, although the difference was not very large.

TABLE IV: Experiment with Stopwords

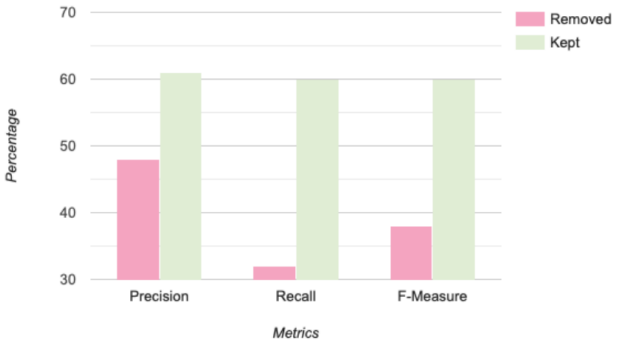

Fig. 8: Stopwords

Fig. 6. 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 routinely removed in training to accelerate calculations. However, eliminating these phrases from a text summary might result in a lower quality final product. Evidence for this was the lack of coherence observed in the final summary after stopwords were removed. Therefore, keeping stopwords for training would be preferable for the BERT model in this particular circumstance.

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.

BERT has proven to perform better than other algorithms in a variety of tasks involving natural language processing. In terms of accuracy and recall, BERT has exceeded well-known algorithms like Support Vector Machines (SVM) and Naive Bayes. In the area of text classification, for instance, BERT has demonstrated outstanding results, beating SVM and Naive Bayes in terms of accuracy rate. Additionally, BERT has shown improved performance in tasks requiring the recognition of named entities and question-answering. Indeed, in benchmarks like the GLUE and Stanford Question Answering Dataset (SQuAD), BERT-based models have broken a number of records. Overall, BERT's remarkable success in a range of NLP tasks may be credited to its ability to understand the links between words and the context in which they are used. Therefore, BERT is preferred over more established techniques like SVM and Naive Bayes in terms of accuracy and recall.

V. Conclusion and Future Work

This study examines the efficacy of the tried-and-true method of extractive summarizing, with an emphasis on applying BERT to anticipate summaries of news stories. The

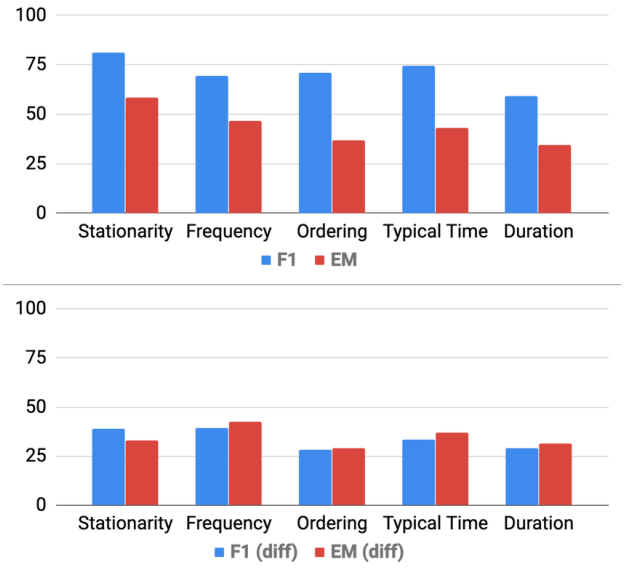

Fig. 7. BERT Performance

experiment's findings demonstrate that, in terms of ROUGE score, precision, recall, and F-measure, BERT performs better than other models. BERT is a better option due to how rapidly data processing occurs as a result of its lightweight architecture. In addition, our research suggests that removing stopwords could harm the accuracy and flow of summaries produced by computers. Additional enhancements to this study include expanding the dataset for more thorough training and applying LSTM and RNNs to produce more abstract summaries similar to human summaries. It is planned to use the universal NLP pipeline developed in this work in the future to summarize texts in other languages, such 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.