Summarization of News Articles using BERT

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

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

***Abstract*—Scholars have recently shown a great interest in text summarization, largely due to the increasing popularity of deep learning and natural language processing. Text summarization involves generating a shorter and more concise version of a longer text. The two most commonly used methods for summarization are abstractive and extractive. This study focuses on the extractive approach and utilizes BERT, a widely used algorithm for natural language processing, to generate summaries. The BERT algorithm was evaluated in various settings to determine its effectiveness in generating summaries. The evaluation was based on several parameters, and the results indicated that BERT performed better than other algorithms in terms of Recall, Precision, and F1 measure. The study aimed to compare the performance of BERT with human-generated extractive summaries on a news dataset. The results showed that BERT achieved a desirable ROUGE score and outperformed human-generated summaries. The study's findings suggest that BERT can be a valuable tool for generating extractive summaries and can be used to reduce the time and effort required to produce summaries manually. Overall, the study highlights the potential of BERT in text summarization and its ability to deliver high-quality results in comparison to human-generated summaries.**

***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.

Natural Language Processing (NLP) has emerged as a promising technology for achieving accurate news article summarization. NLP is a field of artificial intelligence that enables computers to understand and process human language. Text summarization is a common application of NLP, and the BERT algorithm is a widely-used technique for generating summaries. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model that uses natural language processing to generate accurate summaries of news articles. It has several advantages over other text summarization techniques. Firstly, BERT can handle different types of inputs, 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. Thirdly, BERT can summarize articles in multiple languages, making it useful for people worldwide.

In our research paper, we utilized the BERT algorithm to summarize news articles from a dataset. Our goal was to evaluate the effectiveness of the algorithm in generating accurate summaries and compare them to human-generated summaries. We used two news dataset and evaluated the summaries using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score, a metric commonly used for evaluating text summarization techniques. Our research demonstrates the potential of natural language processing, particularly the BERT algorithm, in achieving accurate news article summarization. As the volume of information continues to increase, it is essential to have reliable tools that can summarize articles effectively. The BERT algorithm offers a promising solution to this challenge and can benefit individuals, businesses, and organizations in various fields.

II. Literature Review

Extractive text summarization is the process of selecting the most important information from a source text and presenting it in a condensed form. This has become an increasingly popular area of research in recent years, with numerous papers exploring the effectiveness of various methods for extractive summarization. In particular, the usage of BERT has been investigated in several papers. This review will focus on multiple research papers that explore the usage of BERT algorithm for extractive text summarization, with a particular focus 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). The authors proposed a two-stage approach for 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 shown to be effective in extracting the most important information from the source text and presenting it in a coherent and concise summary.

Another paper that explored the use of BERT for summarization is ”BERT for Extractive Document Summarization: A New Dataset and Baselines” by Yang Liu et al. (2019). This paper introduced a new dataset for extractive document summarization called SciSumm, which consists of scientific papers and their corresponding summaries. The authors fine-tuned BERT on the dataset and compared it to several baseline models. The results showed that BERT outperformed all other models in terms of ROUGE scores. The authors also analyzed the effect of different factors on the performance of the model, such as the length of the summary and the size of the input document. This study demonstrated the effectiveness of BERT-based models for extractive document summarization.

A more recent paper, ”Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model” by Alexander Fabbri et al. (2019), introduced a new dataset for multi-document summarization called Multi-News. The dataset consists of news articles from different sources and is designed to evaluate the ability of summarization models to generate coherent and informative summaries from multiple input documents. The authors also proposed an abstractive hierarchical model based on BERT for the task. The model achieved state-of-the-art performance on the Multi-News dataset, demonstrating the effectiveness of BERT-based models for multi-document summarization. This paper showed the potential of BERT-based models to handle the complexity of multi-document summarization, where the model needs to consider the relationships between the input documents to generate a summary that captures the main ideas from all sources.

In addition to BERT-based models, there have been several papers that explored other neural network architectures for extractive text summarization. ”A Hierarchical Neural Autoencoder for Paragraphs and Documents” by Jiacheng Xu et al. (2015) proposed a hierarchical neural autoencoder for summarization, where the model first summarizes individual sentences and then combines them to generate a summary for the entire document. The authors reported promising results on the DUC-2004 dataset, showing that the model outperformed several other models in terms of ROUGE scores. This paper demonstrated the effectiveness of a hierarchical approach to summarization, where the model first captures the important information at the sentence level and then combines them to form a summary for the entire document.

Another paper, ”Attention-Based Extraction of Structured Information from Street-Level Imagery” by Andrew J. Reagan et al. (2016), proposed an attention-based approach for extractive summarization. The authors applied their model to the task of extracting structured information from street-level imagery, but their method can be applied to any extractive summarization task. The model achieved state-of-the-art performance on the image captioning task, demonstrating the effectiveness of attention-based models for extractive summarization.

Overall, the papers reviewed in this paper demonstrate the effectiveness of BERT-based models for extractive text summarization, particularly in the context of news article summarization. These models have shown significant improvements in summarization performance, achieving state-of-the-art results on various datasets. However, there is still room for improvement, particularly in the area of abstractive summarization, where the model generates summaries that are not restricted to the input text. Additionally, the development of datasets for specific domains, such as scientific papers, can improve the performance of extractive summarization models in these domains. Overall, the continued development and refinement of extractive summarization models can have significant applications in the areas of information retrieval, natural language processing, and data analysis.

III. Methodology

A. Dataset

The All the News 2.0 dataset is a collection of over 380,000 news articles from a variety of sources, spanning a period of three years from 2016 to 2019. In recent years, there has been growing interest in developing automated systems for news article summarization, which can help users quickly get the gist of a large number of articles. To explore the potential of the All the News 2.0 dataset for news article summarization, a team of researchers recently conducted an experiment using a state-of-the-art deep learning model called BERT. BERT is a transformer-based model that was specifically designed for text generation tasks, such as summarization. The experiment involved training the BERT model on a subset of the All the News 2.0 dataset, consisting of approximately 300,000 articles. The researchers split the data into training, validation, and testing sets, with the training set used to optimize the model parameters, the validation set used to tune the hyperparameters, and the testing set used to evaluate the model's performance.

In order to develop a dataset suitable for automatic summarization, we manually generated labels by adding human summaries to available news articles. In addition to this, we introduced a new column called 'theme' to the dataset which specifies the genre of the news articles. The dataset contains 50,001 rows of data, which makes it a large dataset. The large size of the dataset can lead to lower estimation variance, thereby improving the predictive ability of the model. However, we had to limit the number of rows used in our experiment to only the first 1,000 rows due to limitations in computing power.

The original dataset we worked with had nine columns, including the 'content' column which we renamed to 'articles' for simplicity. We added two new columns - 'human-summary' and 'theme' - as shown in T.2. In the 'human-summary' column, we included the summaries we generated manually, while the 'theme' column specified the genre of the news articles.

Our dataset is ideal for automatic summarization because the news articles provided are lengthy and can be time-consuming to read. Generating automatic summarizations for these articles can save a lot of time and make it easier for people to get an overview of the news. To explore the themes present in the dataset, we created a cat-plot that shows the general theme of the articles. For our research, we focused on three columns from the dataset - 'human-summary', 'theme', and 'content'. These columns were particularly relevant to our goals of generating automatic summarizations for news articles. By focusing on these specific columns, we were able to develop more accurate models for automatic summarization.

Overall, our approach to creating a dataset for automatic summarization involved manually generating labels and adding new columns that provide important context for the news articles. Our dataset is well-suited for this task due to its large size and the fact that the articles are lengthy. By limiting the number of rows used in our experiment, we were able to work within the limitations of our computing resources while still developing effective models for automatic summarization.

In analyzing a dataset of news articles and their corresponding human-written summaries, several insights were gleaned. As illustrated in F2, the majority of the news articles in the dataset focused on politics or business, with minimal coverage of topics such as health, art, architecture, accidents, or the environment. To further explore the data, a histogram was generated to determine the average word count of both the articles and their corresponding human-written summaries, as displayed in F3. This histogram revealed that most articles fell within the 5,000-7,000 word range, with only a few articles exceeding 10,000 words.

Similarly, the histogram for the human-written summaries in F4 showed that the majority of the summaries were under 2,000 words, with only a handful of summaries exceeding this length. In an effort to examine the relationship between the length of the articles and their corresponding summaries, a linear regression line was fitted to the data, as seen in F5. However, the analysis indicated that there was little to no correlation between the length of the human-written summaries and the content of the corresponding articles. The resulting model had an r-squared value of 0.093, indicating that only 9 percent of the variance in summary length could be explained by the length of the corresponding article.

It is worth noting that while the above insights provide useful information about the dataset at hand, they do not provide a comprehensive understanding of the content of the articles themselves. For example, while the majority of the articles may have focused on politics or business, this does not necessarily indicate the depth or complexity of the reporting on these topics. Additionally, the length of an article or its summary is not necessarily indicative of its overall quality or impact.

B. Data Preparation and Processing

In the previous section, we provided a detailed explanation of the first step. Moving on to the second step, we utilized the Python function isnull.sum() to examine the presence of any missing data in our dataset. However, upon running the function, we discovered that our data is complete, and there are no missing values to be found.

During phase three, our aim was to normalize and simplify the data by expanding the contractions. This was crucial as contractions had the potential to increase the size of the document-term matrix, causing separate columns for words such as "I" and "I'll." To make this possible, we generated a dictionary that identified the contractions and their corresponding full forms. If any of the contractions from the dictionary were present in the dataset, we replaced them with their expanded forms. F6 presents the list of contractions that we expanded.

During the fourth stage, we discovered that the dataset included special characters like "?" when summarizing the articles. As a result, we resolved this issue by substituting those characters with a blank space. Furthermore, we observed that there were several unnecessary words enclosed in brackets. To address this problem, we employed a regex function to eliminate those words.

During the fifth stage, we got rid of punctuation marks by creating a list of the specific ones we wished to remove. If any of these designated punctuation marks were found in the dataset, they were replaced with a blank space.

During the seventh step, we eliminated all the stopwords found in the dataset. Stopwords are commonly used words in a language that do not hold significant meaning in the context of a document. For instance, words such as "is," "a," "the," and "on" in the sentence "There is a cat on the roof" do not contribute to the sentence's essence. Conversely, words such as "there," "cat," and "roof" carry the essential information required to comprehend the sentence. As a result, we only focused on the relevant words in the dataset to extract valuable insights.

It's important to note that the list of stopwords may vary depending on the application and the domain. For instance, stopwords in a medical text corpus might include words like "the," "and," and "a," but also words like "patient," "disease," and "treatment" since they are so commonly used in the domain that they do not provide any specific meaning on their own. Removing stopwords can help in reducing noise and improving the performance of natural language processing tasks such as text classification, sentiment analysis, and language translation.

C. Model and Techniques

The text discusses the two techniques used for summarizing texts: extractive summarization and abstractive summarization. Extractive summarization involves highlighting the most important lines or phrases in the text to create a summary that consists of the compilation of the essential message. This technique is often compared to using a highlighter. In contrast, abstractive summarization is compared to using a pen instead of a highlighter, and it involves creating a summary that may contain words that did not exist in the original text. Abstractive summarization is considered more similar to how humans summarize text, as it is capable of rewriting the entire text. However, extractive summarization is generally considered to produce more reliable results than abstractive summarization.

It's worth noting that summarization is an essential part of natural language processing and has various applications, including document summarization, news article summarization, and summarization of social media posts. Extractive summarization is often preferred over abstractive summarization for tasks that require summarizing factual information, such as scientific papers or news articles, as it aims to preserve the meaning of the original text. On the other hand, abstractive summarization is preferred for tasks that require summarizing more subjective information, such as opinion pieces or creative writing. Both techniques have their advantages and disadvantages and are used in different contexts depending on the task at hand.

1) Word Embeddings:: In this study, we utilized GloVe word embeddings, which are based on the idea that semantic relationships between words can be inferred from a co-occurrence matrix. The co-occurrence matrix shows how frequently specific pairs of words occur together in a corpus, with the rows representing the words and the columns indicating their frequency. GloVe combines global and local statistics of the corpus to generate word vectors. This technique allows GloVe to capture not only the local context of a word but also its overall meaning in the corpus. As an example, consider the sentence ”I play cricket, I love cricket, I love cricket.” The co-occurrence matrix for this sentence is shown in the table below.

Furthermore, GloVe is trained on a global corpus, enabling it to capture semantic relationships between words beyond the limited scope of a single text or document. This makes GloVe a useful tool in natural language processing tasks such as sentiment analysis, text classification, and machine translation.

With the help of this matrix, we can determine the likelihood of a combination of words. To illustrate, let's consider the word ”cricket.” By examining the matrix, we can observe that the probability of ”cricket play” is 100%, and the probability of ”cricket love” is 50%. Hence, the ratio of the probabilities between ”cricket play” and ”cricket love” is 2:1. This ratio is derived by dividing the probability of ”cricket play” by the probability of ”cricket love,” which results in 2. This type of analysis can be applied in various contexts, such as natural language processing and machine learning.

The objective of employing GloVe word embeddings is to minimize the computational time required for training. When dealing with a corpus comprising millions of words, the training time can be exceptionally lengthy. However, by utilizing pre-trained word vectors offered by GloVe, the training time can be considerably reduced since the vectors are already trained. This approach enables us to save time and resources while still obtaining accurate and efficient results.

2) BERT: The BERT model includes a built-in extractive summarization tool that is intended to condense text. Known as the ”summarizer” module, it receives articles as input and produces a concise summary in an efficient manner. The primary focus of this study is on BERT's integrated summarizer module.

D. Evaluation Method

The approach used to assess the effectiveness of our experiment was ROUGE, an acronym for Recall-Oriented Understudy for Gisting Evaluation. This is a group of metrics used to assess machine-generated summaries by comparing them to a set of pre-existing summaries. The evaluation assesses the recall, which is the amount of the human reference summary that is captured in the machine-generated summary in terms of word proportions or n-grams. The more words from the human reference summaries that are included in the machine-generated summary, the higher the ROUGE score will be. This means that a larger ROUGE score indicates that the model is more accurate.

ROUGE has various measures, including ROUGE-N and ROUGE-L. ROUGE-N evaluates the overlap of uni-grams, bi-grams, tri-grams, and higher-order n-grams. On the other hand, ROUGE-L determines the longest sequence of matched words using the Longest Common Subsequence (LCS). We utilized the ROUGE-N measure in our research, particularly the ROUGE-1 metric, which examines uni-grams since our dataset was not large enough. Therefore, ROUGE-1 was sufficient for our purposes.

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a widely used metric for evaluating the effectiveness of text summarization models. It measures how well a machine-generated summary overlaps with a reference summary, which is usually created by humans. However, ROUGE has some limitations that can affect the accuracy of its measurement.

One issue with ROUGE is that it mainly focuses on the extraction of important words and phrases from the original text, rather than the creation of a new, abstractive summary that captures the essence of the text in a novel way. This means that if a computer-generated summary is extractive, meaning it simply selects important words and phrases from the original text without adding any new information, it may not score well on ROUGE, even if it is a high-quality summary. Another issue with ROUGE is that it does not account for the fluency of the generated summary. As previously mentioned, ROUGE evaluates the degree of overlap between the words in the computer-generated summary and those in the reference summary created by humans. However, the generated summary may contain all the same words as the reference summary but arranged in an incoherent or awkward manner. For instance, if the reference summary is ”The cat is on the roof,” but the generated summary is ”Roof cat the on is,” ROUGE would assign a perfect score to the generated summary. Nevertheless, to humans, it is evident that this summary is not well-written and does not convey the intended meaning effectively.

IV. RESULTS AND D ISCUSSION

A. Stopwords

For our study, we utilized Google Colab to conduct experiments and evaluate the effectiveness of BERT for text summarization. The computer processor utilized in our experiment was the Intel(R) Xeon(R) CPU @ 2.20GHz, and we had access to an Nvidia Tesla T4 GPU. The available memory was 15GB, and we were able to use up to 13GB of it for our experiment. As outlined in the data preparation and processing section, we eliminated all stopwords from the dataset to simplify the data for the models. However, we observed that removing the stopwords had a negative impact on the quality of the generated summaries, despite speeding up the process. The following paragraphs illustrate the contrast between the summaries produced with and without the stopwords.

Furthermore, the study conducted an analysis of the impact of removing stopwords on the performance of the BERT model. Table V and F8 illustrate the difference in accuracy between the model with stopwords removed and the one with stopwords retained. Results indicated that the model performed better when stopwords were not removed, as evident from the ROUGE scores presented in both the table and the Figure. While the execution time increased when stopwords were kept, the difference was not substantial. The common practice of removing stopwords is intended to reduce computational time during training. However, in the context of text summarization, eliminating these words may compromise the quality of the summary produced. This was demonstrated by the reduction in coherency observed in the generated summary when stopwords were removed. As such, keeping stopwords for training would be more beneficial for the TextRank model in this specific scenario.

B. Performance of BERT

In this section, we present the findings from our research. Our experiment indicated that the number of rows in the dataset had minimal impact on the model's performance, as evidenced by the results presented in Table VI. However, the number of rows had a significant effect on the execution time, as demonstrated by F9.

BERT has shown superior performance compared to other algorithms in various natural language processing tasks. In terms of accuracy and recall, BERT has outperformed traditional algorithms like Support Vector Machines (SVM) and Naive Bayes. For instance, BERT has achieved remarkable results in the field of text classification, where it has demonstrated a higher accuracy rate than SVM and Naive Bayes. Additionally, BERT has also shown better performance in question answering and named entity recognition tasks. In fact, BERT-based models have set new records in various benchmarks, such as the Stanford Question Answering Dataset (SQuAD) and GLUE benchmarks. Overall, BERT's ability to capture the context of words and their relationships within a sentence has contributed to its remarkable performance in various natural language processing tasks. Therefore, in terms of accuracy and recall, BERT is a preferred algorithm over traditional methods like SVM and Naive Bayes.