Journal of Theoretical and Applied Information Technology

31st October 2023. Vol.101. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPEED UP THE DEEP BIDIRECTIONAL TRANSFORMERS WITH FEATURE SELECTION FOR SUMMARIZING MEDICAL PAPERS

ALSHIMAA.M.IBRAHIM, MOSTAFA MAHMOUD AREF, MARCO ALFONSE

Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

E-mail: alshimaa.mohamed@cis.asu.edu.eg, mostafa.araf@cis.asu.edu.eg, marco_alfonse@cis.asu.edu.eg

ABSTRACT

Medical papers are being widely published currently, especially after the Coronavirus disease (COVID-19) pandemic. The time required to manually summarize medical papers can be decreased by applying text summarization approaches. It is now common practice to overcome medical text summarization challenges using pre-trained models such as the Bidirectional Encoder Representations from Transformers (BERT)-base model. This paper presents a new system for summarizing medical papers based on deep learning techniques. In this system, we combine the x^2-Statistic (CHI-square) feature selection technique with a token classification such as Part-of-Speech (POS) tagging and use the feature selection output as input to the pre-training BERT-base model, then apply clustering algorithms for the sentence selection process. Our main contribution is that our model obtained high speed and accuracy compared to previous summarization methods. We performed our comprehensive experiment on the public corpus that was randomly selected from BioMed Central (BMC). In comparison to other models that need a lot of training time, our model's output has high performance and is less complex. We used the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric for an evaluation process. The model results are ROUGE-1 = 0.7611, ROUGE-2 = 0.3205, and ROUGE-L=0.4544.

Keywords: *Text Summarization, Machine Learning, Deep Learning, Natural Language Processing, Feature Selection.*

1. INTRODUCTION

The number of papers published in the medical field is growing every day and applying text summarization techniques can cut down the time needed to personally summarize medical papers. Text summarization has now been widely examined and put to use in numerous contexts and practical applications, especially in the medical field. Examples include summaries of COVID-19 research publications [1], voice-based helpers, search engines [2], etc. The two main approaches in text summarization are extractive and abstractive. The extractive summary comes first and aims to extract and concatenate key passages from the source text. The abstractive technique focuses on creating fresh summaries that paraphrase the underlying text.

The deep bidirectional transformers allow the model to learn information from right to left, and from left to right. Comparing the bidirectional models to right-to-left and left-to-right models reveals that bidirectional models are significantly more effective. The BERT model is employed as the pretraining of the deep bidirectional transformers for language understanding. Traditional feature selection techniques like TF-IDF [3], TextRank [4], Mutual Information (MI) [5], and x^2-Statistic (CHI-square) [6] have been employed to enhance BERT's performance on large texts. Based on an experiment in [7], chi-square was employed as the feature selection to enhance performance and condense the text while preserving its essential information. The condensed text was then entered into the BERT-base model. The pre-training model BERT-base [8] was subsequently introduced, and it has demonstrated outstanding performance in a multitude of Natural Language Processing (NLP) tasks. The multi-layer transformers that form the network architecture of BERT are instituted on multi-head self-attention [9]. Transformers need a lot of time and GPU memory to process long sequences because of their $O(n^2)$ computational and spatial complexity, where

n is the length of the input sequence. As a result, BERT needs a lot of GPU memory to process long texts. In this paper, we used the extractive approach to summarize text by proposing a new system for summarizing medical papers based on deep learning techniques by combining the BERT-base model with the x^2-Statistic (CHI-square) feature selection technique and POS tagging that speed up the execution time of summarization and obtained high accuracy compared to previous summarization methods.

The following describes the structure of this paper: section 2 presents the related work. Section 3 provides the methodology used to implement our model. Section 4 presents the evaluation of our study. Section 5 shows the results and discussion. Section 6 presents the conclusion and future work.

2. RELATED WORK

Recently, extensive studies based on machine learning and deep learning for text summarization have been presented. This section provides related work that employs a multitude of summarization techniques. Instead of sentence production for text summarization, most of the research focuses on sentence extraction. The most frequently employed summarization approach depends on statistical aspects of the text that generate extractive summaries.

Wang K et al [7] proved that merging feature selection methods, especially the CHI-square method with the BERT model obtained high accuracy to fix the long text classification issue. The authors used a real-world and public dataset from China Telecom that was divided into 12133 documents for training and 2599 for testing. They used accuracy, Precision, Recall, F1-score, and Hamming-loss metrics for the evaluation process. The result for BERT+CHI is accuracy = 0.611, precision= 0.589, recall= 0.611, F1-score= 0.592, hamming-loss= 0.389.

Kieuvongngam et al [10] applied token classification such as part of speech tagging (POS) with pre-trained BERT and an Open-source Artificial Intelligence Generative Pre-trained Transformer - 2 (OpenAI GPT-2) models. The COVID-19 open research dataset was used that consists of 31246 articles for training and 3572 articles for testing. They used ROUGE metrics to evaluate the experiment. The result of ROUGE-1 is 0.392, ROUGE-2 is 0.159, and ROUGE-L is 0.305.

Jabri et al [11] used feature selection techniques such as the CHI-square method to summarize text and then applied a Support vector machine (SVM) to classify the summarized text. They proved that using classification with chi-square achieved high accuracy rather than chi-square only. Eight hundred Arabic text documents make up the corpus. It is a portion of the 60913-document corpus that was gathered from a multitude of websites. The authors used precision and recall for evaluation. The model result is obtained as follows, precision=0.825 and recall= 0.808.

Moradi et al [12] improved medical text summarization outcomes by combining a deep bidirectional language (BERT) model and a clustering algorithm without the need for computationally demanding knowledge bases or domain-specific pre-training. They generated a new corpus by randomly selecting 4000 articles from BioMed Central (BMC). The authors utilized the ROUGE metric for the evaluation process and the result is ROUGE-1=0.7504 and ROUGE-2=0.3312.

Moradi [13] presented an autonomous summarization system by combining a clustering algorithm with an itemset mining algorithm that yields the best results. The outputs showed that a topic-based sentence clustering method increased the summary's meaningful material while reducing its unnecessary details. They utilized a corpus of four hundred scholarly biomedical papers from the BMC corpus as a single document. The ROUGE metric is utilized to evaluate the result. The ROUGE-1 value is 0.7345 and ROUGE-2 is 0.3187.

Moradi, and Nasser Ghadiri [14] created a summarization system by utilizing a Unified Medical Language System (UMLS) and then applied different feature selection approaches to identify the essential sentences in the documents. The authors created the dataset used in the experiment by randomly selecting 400 biomedical papers from the BMC corpus. The authors used the ROUGE metric for evaluation. The result of ROUGE-1 is 0.7288, and ROUGE-2 is 0.3143.

Saggion [15] created a summarization tool called Scalable Understanding of Multilingual Media (SUMMA) based on the General Architecture for Text Engineering (GATE) platform. Moradi et al [12] used the SUMMA system with the BioMed Central corpus which consists of 4000 articles and utilized ROUGE in the SUMMA summarizer. The

result of ROUGE-1 is 0.7098, and ROUGE-2 is 0.3022.

Text Analytics for Law Enforcement Agencies (TexLexAn) [16] is a system that enhances large amounts of text data in different domains based on a linear classifier, fuzzy logic, and case-based reasoning. Moradi et al [12] utilized the TexLexAn system with the BMC dataset which consists of 4000 articles and apply the ROUGE in the TexLexAn summarizer. The result of ROUGE-1 is 0.6982, and ROUGE-2 is 0.2979.

According to the previous work, we found that multiple feature selection strategies, including TF-IDF, TextRank, CHI-square, and Mutual Information, were used by Wang et al. [7] with the BERT model. They demonstrated that, in comparison to previous feature selection strategies, employing the BERT model with CHI-square feature selection increases the performance of text summaries, but they did not make use of more recent pre-processing techniques. The BERT and OpenAI GPT-2 pre-trained models were employed by Kieuvongngam et al. [10]. The study's computational capacity is constrained by the use of the GPU rather than the DistilGPT2 version. A model by Jabri et al. [11] that combines SVM and the CHI-square feature selection technique produced accuracy levels that were superior to those obtained by CHI-square alone. The study has a flaw because the summary takes longer to execute. The main limitation of Moradi et al. work [12] is the rareness of available datasets with their gold summarization. Moradi [13] has a weakness in that it could be unable to capture the overall structure of a document, which can generate an unsuitable summary. Moradi, and Nasser Ghadiri [14] applied the generated model in a small data set. For less-resourced languages, Saggion's summarization tool [15] performed less accurately and with less precision due to limited training data. TexLexAn [16], did not utilize modern pre-processing methods to improve the outcome and solve the problem of data quality and noise.

3. METHODOLOGY

Our text summarization process is divided into many steps: pre-processing, feature selection, feature extraction using a BERT-base model, sentence clustering, and selection. Our architecture is depicted in Figure 1. Each step is thoroughly explained in the subsections that follow.

3.1. PRE-PROCESSING

Preprocessing enhances outcomes, minimizes computations, and boosts speed and accuracy. Several preprocessing methods are employed in this article: tokenizers**,** stop words, and POS tagging.

- **TOKENIZERS**

  Text is divided into tokens. Sentences, regex tokenizers, and words are the three main tokens found in tokenizers. Only words and sentence tokenizers are utilized in our experiment [17].

- **STOP WORDS**

  One of the preprocessing techniques that are most frequently utilized across many NLP applications is stop word removal. The simple idea is to exclude words that appear frequently throughout all the corpus's documents. Pronouns and articles are typically categorized as stop words. These words are not extremely discriminative.

- **POS TAGGING**

  POS tagging is a further crucial technique that needs to be used. It is the process of assigning various parts of speech to specific words inside a sentence. This is done at the token level, as opposed to phrase matching, which is done at the sentence or multi-word level [18].

3.2. FEATURE SELECTION

It is a procedure where you can automatically choose the corpus features that contribute the prediction variable or the most to the output that interests you. The advantages of feature selection before data modeling are:

1. Avoid overfitting as less duplicated data improves model performance and reduces the chance that decisions will be based on noise.
2. Reduces training time.

The x^2-Statistic (CHI-Square) feature selection is a popular technique for choosing features from text data. It is a measurement of the discrepancy between the frequency of outcomes of a collection of events or variables that are observed and those that are

predicted. The chi-Square formula is represented in Equation (1).

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where c is degrees of freedom, O is the observed value(s), E is an expected value(s)

3.3. BERT MODEL

The Bidirectional Encoder Representations from the Transformers (BERT) model is used to pretraining the deep bidirectional transformers from the unsupervised text by working together on right and left context in whole layers for language understanding. The BERT model consists of 2 steps; pretraining and finetuning. Throughout pre-training, the model is trained on unsupervised data. During fine-tuning, the BERT model is started with the pre-trained parameters, then the parameters are fine-tuned using supervised data. BERT is a multi-layer bidirectional Transformer encoder depending on Transformer. The transformer is created by stacking numerous encoders and decoders. Segment, token, and position embeddings are the three feature embeddings combined in the input of BERT. Two sentences can be distinguished using segment embeddings. Each word is changed into a fixed-dimensional vector using token embedding. The position information of the word is encoded into a feature vector through position embedding. There are two types of BERT; BERT-base (Layers = 12, Hidden size =768, Self-attention heads =12, Total Parameters=110 Million), and BERT-large (Layers = 24, Hidden size =1024, Self-attention heads =16, Total Parameters=340 Million).

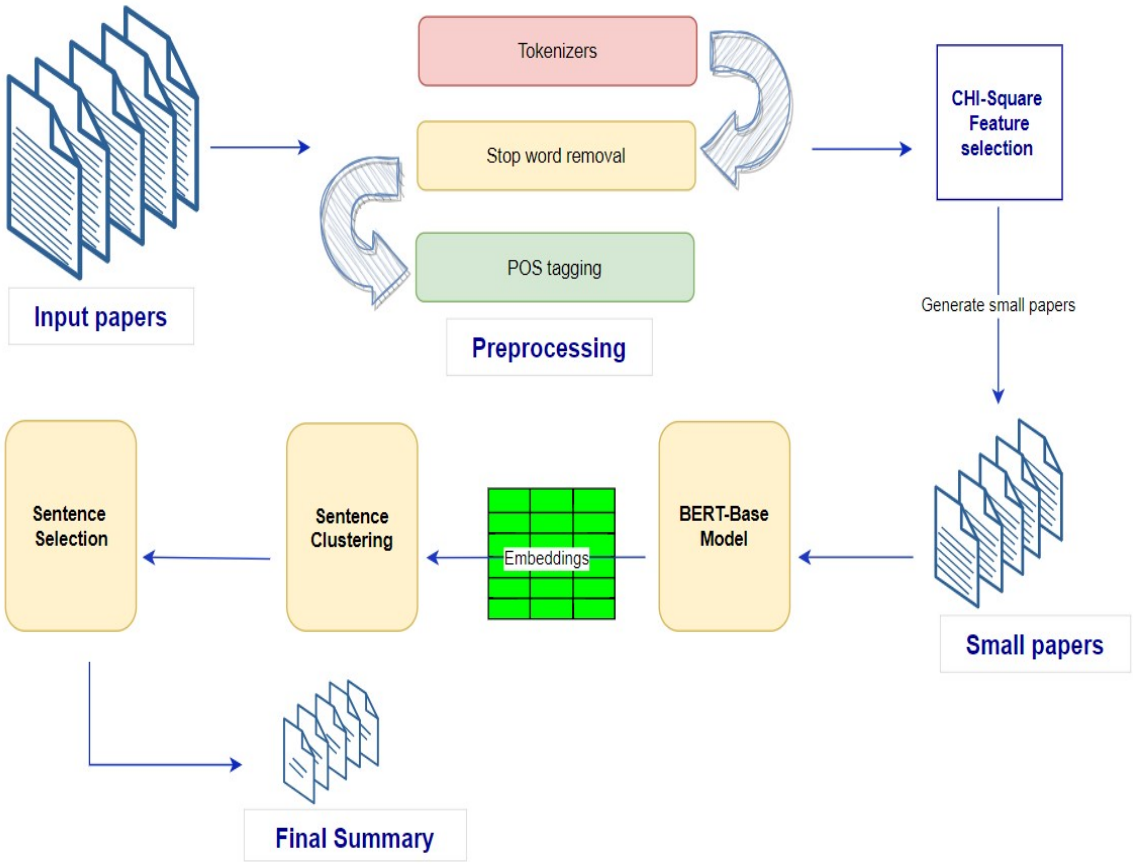

Figure 1. The proposed summarization model architecture

3.4. SENTENCE CLUSTERING AND SELECTION

We utilized the closeness of sentences in the vector space to determine how much of their content is similar. It is thought that sentences with neighboring vectors have some context. The summarizer employs clustering to group identical sentences according to the distance between their vector space representations. Through a process known as agglomerative hierarchical clustering, the summarizer creates clusters of sentences. We used a sentence clustering algorithm from [12]. The relatedness of phrases within a clustering algorithm can be evaluated using a multitude of measures. Euclidean distance and cosine similarity are often used to determine how strongly connected objects are in a vector space. These two measures cover different facets of vectors. In contrast to cosine similarity, which deals with the direction of the vectors or the angle between vectors, euclidean distance is concerned with the magnitude of the vectors in multiple dimensions.

Euclidean distance is represented in Equation (2):

$$\text{Euclidean distance (A, B)} = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2} \quad (2)$$

where B=$\{b_1, b_2, \ldots, b_N\}$ and A=$\{a_1, a, \ldots, a_N\}$ are two vectors matching the two sentences given.

Cosine similarity is represented in Equation (3):

$$\text{Cosine similarity (A,B)} = \frac{A.B}{||A||\,||B||} \quad (3)$$

where B=$\{b_1, b_2, \ldots, b_N\}$ and A=$\{a, a_2, \ldots, a_N\}$ are two vectors matching the two sentences given.

The sentence selection process chooses instructive sentences from each cluster to create the summary. Each cluster's size can serve as a sign of its significance. Equation (4) is used by the summarizer to specify the number of sentences that must be chosen from each cluster.

$$N_i = N \frac{|C_i|}{|D|} \quad (4)$$

where N_i is how many sentences the summarizer selects from cluster Ci for use in the summary, $|C_i|$ is the size of cluster Ci, N is the total number of sentences that must be chosen for inclusion in the summary, and |D| is the total number of sentences.

The summarizer assigns a score to each sentence inside each cluster based on the informativeness scores that Equation (4) calculated. The summarizer then groups the sentences, arranges them according to the sequence in which they appeared in the initial document, and creates the output summary.

4. EVALUATION

4.1. DATASET

We conducted our research on an open-access publisher dataset selected from the BioMed Central (BMC) dataset to show the efficacy of our approach. Moradi et al [12] produced a corpus from BioMed Central (BMC) by using the abstracts as model summaries with 1000 articles for training, and 3000 articles for testing then they added them to their new corpus. We used the same dataset for evaluating our model. As model summaries, abstracts are used. This size of the evaluation and development corpora is sufficient to guarantee that the results are statistically significant, as Lin [19] showed.

4.2. EVALUATION METRICS

We used the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric for the evaluation process [20] that automatically assesses the quality of a summary by comparing it to other gold summaries. It includes various measures, including ROUGE-N. We employ ROUGE-1, ROUGE-2, and ROUGE-L in our evaluation. ROUGE-1 assesses the content overlap in terms of common unigrams, whereas ROUGE-2 takes shared bigrams. ROUGE-N counts the number of word pairs, word sequences, and n-grams that overlap between the model-produced summary and the gold summaries created by the reviewers [20]. Equation (5) provides the ROUGE-N formula.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (5)$$

where $gram_n$ is a common n-grams that is included in an elect summary, n is a representation of the n-gram length such as ROUGE-1 and ROUGE-2, and $Count_{match}(gram_n)$ is a group of gold summaries.

The similarity of the two sequences is represented by Rouge-L. It is based on a Longest Common Subsequence (LCS). The formula for the ROUGE-L metric is represented in Equation (6).

$$\text{ROUGE-L} = \frac{(1+\beta^2)\,Recall * Precision}{Recall + \beta^2 * Precision} \quad (6)$$

where β set the value of recall and precision according to each other and set to a high value.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Equation (7) and Equation (8) represent Precision, and Recall respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

where TP is a True Positive, FP is a False Positive and FN is a False Negative.

4.3. LEARNING ENVIRONMENT

The hardware configuration used in the experiments is shown in Table 1.

Table 1. Hardware configuration

Hardware	Configuration
System	Windows Server 2019
RAM	8GB
CPU	Intel(R) Xeon(R) Silver 4216 CPU @2.10GHz

Our summarization model takes 6 days to finish summarization of 3000 articles while utilizing the same environment shown in Table 1 to run the BERT-based summarizer of [12], it takes more than 60 days to summarize the same 3000 articles. So, it is clear that our model is more efficient according to the required summarization time.

5. RESULTS AND DISCUSSION

We used the optimal settings recommended by [12] to run our model. The compression rate was set to 0.3 in all experiments. As a result, the size of the summary does not exceed 30 percent of the original summary [21]. The number of clusters in the sentence clustering algorithm is indicated by parameter K. Various K parameter values (between 2 and 12) were tested. The ROUGE-1, ROUGE-2, and ROUGE-L results were obtained by applying the BERT-base model to the full-text publications using the feature selection technique of CHI-square, Euclidean distance, cosine similarity, and summarized ones. We compared our model with the results of a related study [12] that applied the Euclidean distance and the cosine similarity as clustering algorithms and utilized different versions of the BERT model.

The ROUGE-1 values of our model compared to different versions of the BERT-based summarizer in [12]; BERT-Base, BERT-Large, BioBERT-PMC, BioBERT-PubMed, and BioBERT- PubMed + PMC using Euclidean distance as the measure of the clustering algorithm is shown in Table 2. The top score is highlighted in bold text.

Table 2. The ROUGE-1 values of our model compared to different versions of the BERT-based summarizer in [12] using Euclidean distance as the measure of the clustering algorithm

K	Our model	BERT-Base	BERT-large	BioBERT - PMC	BioBERT - PubMed	BioBERT-PubMed + PMC
2	**0.7611**	0.7221	0.7434	0.7243	0.7269	0.7369
3	0.7569	**0.7291**	0.7457	**0.7308**	**0.7361**	**0.7429**
4	0.7544	0.7224	**0.7507**	0.7299	0.7354	0.7399
5	0.7534	0.7205	0.7467	0.7272	0.7293	0.7398
6	0.7516	0.7199	0.7415	0.7272	0.7276	0.7352
7	0.7498	0.7157	0.7366	0.7187	0.7226	0.7313
8	0.7481	0.7179	0.7334	0.7194	0.7198	0.7272
9	0.7457	0.7146	0.7291	0.7194	0.7174	0.7273
10	0.7430	0.7127	0.7284	0.7186	0.7162	0.7196
11	0.7401	0.7063	0.7257	0.7186	0.7113	0.7164
12	0.7374	0.7034	0.7203	0.7094	0.7087	0.7117

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The ROUGE-2 values of our model compared to different versions of the BERT-based summarizer in [12]; BERT-Base, BERT-Large, BioBERT-PMC, BioBERT-PubMed, and BioBERT- PubMed + PMC using Euclidean distance as the measure of the clustering algorithm is shown in Table 3. The top score is highlighted in bold text.

Table 3. The ROUGE-2 values of our model compared to different versions of the BERT-based summarizer in [12] using Euclidean distance as the measure of the clustering algorithm

K	Our model	BERT-Base	BERT-large	BioBERT- PMC	BioBERT-PubMed	BioBERT-PubMed + PMC
2	**0.3205**	0.3087	0.3264	0.3094	0.3122	0.3195
3	0.3184	**0.3133**	0.3285	0.3172	0.3186	**0.3265**
4	0.3134	0.3107	**0.3329**	**0.3189**	**0.3187**	0.3234
5	0.3138	0.3114	0.3302	0.3138	0.3183	0.3229
6	0.3109	0.3099	0.3249	0.3134	0.3146	0.3199
7	0.3093	0.3075	0.3208	0.3097	0.3111	0.3170
8	0.3085	0.3079	0.3183	0.3089	0.3074	0.3122
9	0.3061	0.3084	0.3173	0.3099	0.3062	0.3087
10	0.3057	0.3054	0.3137	0.3102	0.3036	0.3080
11	0.3021	0.2990	0.3089	0.3161	0.2992	0.3027
12	0.3010	0.2968	0.3101	0.3088	0.2995	0.3006

The ROUGE-1 values of our model compared to different versions of the BERT-based summarizer in [12]; BERT-Base, BERT-Large, BioBERT-PMC, BioBERT-PubMed, and BioBERT- PubMed + PMC using cosine similarity as the measure of the clustering algorithm is shown in Table 4. The top score is highlighted in bold text.

Table 4. The ROUGE-1 values of our model compared to different versions of the BERT-based summarizer in [12] using cosine similarity as the measure of the clustering algorithm

K	Our model	BERT-Base	BERT-large	BioBERT - PMC	BioBERT-PubMed	BioBERT - PubMed + PMC
2	0.7549	0.7196	0.7328	0.7242	0.7177	0.7285
3	0.7559	0.7169	**0.7377**	0.7249	0.7224	**0.7328**
4	0.7559	**0.7212**	0.7362	**0.7272**	0.7268	0.7278
5	0.7555	0.7152	0.7361	0.7212	**0.7298**	0.7295
6	0.7557	0.7136	0.7299	0.7171	0.7261	0.7272
7	0.7536	0.7107	0.7259	0.7173	0.7221	0.7224
8	**0.7561**	0.7071	0.7231	0.7176	0.7207	0.7199
9	0.7549	0.7037	0.7194	0.7119	0.7170	0.7182
10	0.7544	0.6989	0.7173	0.7073	0.7143	0.7158
11	0.7560	0.6953	0.7146	0.7035	0.7080	0.7126
12	**0.7561**	0.6908	0.7142	0.6995	0.7033	0.7106

The ROUGE-2 values of our model compared to different versions of the BERT-based summarizer in [12]; BERT-Base, BERT-Large, BioBERT-PMC, BioBERT-PubMed, and BioBERT- PubMed + PMC using cosine similarity as the measure of the clustering algorithm is shown in Table 5. The top score is highlighted in bold text.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5. The ROUGE-2 values of our model compared to different versions of the BERT-based summarizer in [12] using cosine similarity as the measure of the clustering algorithm

K	Our model	BERT-Base	BERT-large	BioBERT-PMC	BioBERT-PubMed	BioBERT-PubMed + PMC
2	0.3158	0.3092	0.3224	0.3117	0.3095	0.3163
3	**0.3168**	0.3102	**0.3275**	**0.3131**	0.3089	**0.3204**
4	0.3166	**0.3107**	0.3249	0.3107	**0.3184**	0.3202
5	0.3153	0.3068	0.3259	0.3082	0.3165	0.3199
6	0.3151	0.3026	0.3205	0.3071	0.3157	0.3160
7	0.3115	0.2984	0.3162	0.3008	0.3126	0.3136
8	**0.3168**	0.2988	0.3127	0.3049	0.3102	0.3135
9	0.3143	0.2968	0.3094	0.3001	0.3072	0.3099
10	0.3144	0.2917	0.3068	0.2965	0.3056	0.3074
11	0.3162	0.2905	0.3046	0.2954	0.2986	0.3069
12	0.3146	0.2879	0.3018	0.2882	0.2967	0.3034

The values of the ROUGE-L metric of our model using Euclidean distance and cosine similarity as the measure of the clustering algorithm and applying various K are shown in Table 6.

Table 6. The values of ROUGE-L of our model using Euclidean distance and cosine similarity

K	Using Euclidean distance	Using cosine similarity
2	**0.4544**	0.4493
3	0.4515	**0.4499**
4	0.4475	0.4495
5	0.4476	0.4489
6	0.4451	0.4489
7	0.4428	0.4465
8	0.4423	0.4494
9	0.4403	0.4483
10	0.4387	0.4474
11	0.4360	0.4493
12	0.4337	0.4487

The comparison between our model and other summarization models using ROUGE-1 and ROUGE-2 metrics is shown in Table 7. When we utilized the BERT-base in our summarizer, it obtained a high result against all other comparison methods in terms of ROUGE-1 and slightly less in terms of ROUGE-2 compared to BERT-based summarizer (BERT-large), BERT-based summarizer (BioBERT-pubmed+pmc).

Table 7. The comparison between our model and other summarization models

Summarizer	ROUGE-1	ROUGE-2
Our Summarizer (BERT-base)	**0.7611**	**0.3205**
BERT-based summarizer (BERT-large) [12]	0.7504	**0.3312**
BERT-based summarizer (BioBERT- PubMed +PMC) [22]	0.7411	**0.3228**
BERT-based summarizer (BioBERT-PubMed) [22]	0.7376	0.3203
CIBS biomedical summarizer [13]	0.7345	0.3187
BERT-based summarizer (BioBERT-PMC) [22]	0.7309	0.3164
Bayesian biomedical summarizer [14]	0.7288	0.3143
BERT-based summarizer (BERT-base) [12]	0.7257	0.3110
SUMMA [15]	0.7098	0.3022
TexLexAn [16]	0.6982	0.2979

The results demonstrate that Euclidean distance is more appropriate and faster than cosine similarity to evaluate the relatedness between phrases. This shows that for assessing the relationship between phrases in this kind of contextualized representation, the magnitude of vectors may be more valuable than their direction. We used BERT-base in our experiment, and it was combined with feature selection. Our model achieves higher results and less execution time compared to the original BERT-base utilized in the BERT-based summarizer in [12] in terms of both ROUGE-1 by 0.0354 and ROUGE-2 by 0.0095 evaluation metrics but when compared to the BERT-large and BioBERT-PubMED+PMC models utilized in the BERT-based summarizer in [12], our model achieves high results in ROUGE-1 by 0.0107 and slightly less in ROUGE-2 by 0.0107. Based on the previous

experiments, the BERT-Large and BioBERT-PubMED+PMC models are more efficient than the BERT-base model but need more powerful GPUs [9]. We applied our study for extractive summarization only, we will apply our model to abstractive summarization. Previous research indicates that when small values, like 2, are used for parameter K, the clusters may not properly divide words according to the context they share. Since the most pertinent lines from larger clusters still had the greatest coherence ratings, in this situation, the sentences of smaller clusters are consolidated into larger clusters and loss to be featured in the summary. As a result, the summary may contain redundant information, which would reduce its informative content.

In our model, the performance gets better when the coefficient K is in the interval [2,6] in the Euclidean distance, and between 3 and 6 or being [8,11,12] in cosine similarity. The performance begins to decrease gradually when the coefficient K is higher than 6 in Euclidean distance. On the opposite side, the ROUGE-1 increases when the K value increases in cosine similarity. After integrating the feature selection method, it selects the words in the most frequently used sentences in the document, allowing the model to focus on the sentences that are more important than others that are less important in the document. In the end, the results demonstrate that the number of clusters is large enough to allow informative sentences within smaller clusters to appear in the summary and small enough to avoid the formation of irrelevant sentences in the summary. This shows that parameter K can be crucial in balancing summaries' informativeness against information redundancy.

Figure 2 provides an example of our model summary and its gold summary presented in [23]. A quicker and more thorough understanding of the article is possible by combining this concise summary with the abstract.

6. CONCLUSION AND FUTURE WORK

In this paper, we provided a deep learning-based model for the extractive summarization of medical papers because the number of publications is growing every day in the medical field, and applying text summarization techniques can decrease the time needed to manually summarize medical papers. The main contribution of this study is to demonstrate how contextualized embeddings created by the BERT-Base model speed up the summary process and enhance medical text summarization performance. Also, the results show that when assessing the relatedness between phrases, Euclidean distance is more accurate and quicker than cosine similarity. This demonstrates that, in this type of contextualized representation, the magnitude of vectors may be more valuable than their direction. To better identify the most informative sentences and speed up our model execution, we combined part-of-speech (POS) tagging, feature selection, the BERT-base model, and sentence clustering techniques. We utilized the standard evaluation metrics ROUGE-1, ROUGE-2, and ROUGE-L in comparing our model with other models. The experiments demonstrated that our approach generates better outcomes than other summarization models while being less complex and faster than other methods. Our model results are ROUGE-1 = 0.7611, ROUGE-2 = 0.3205, and ROUGE-L=0.4544.

In the future, we are going to apply our model to different fields other than medical papers and evaluate its performance. Also, we will apply our model to different datasets in different fields and domains. In addition to applying it to abstractive summarization.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Output summary	Some studies have discussed the incidence of and mortality rates differences between Asian Americans and other ethnic groups, but few examined disparities among Asian subgroup populations (ChienNone, 2005; WhiteNone, 2010). The purpose of this analysis was to determine whether there are significant differences in presentation, clinicopathologic features, treatment, and survival rates between non-Hispanic white (NHW) and Asian; and between Asian subgroups in patients with CRC. In addition, lifestyle factors and screening prevalence were analyzed to determine whether any disparities exist between the Asian subgroups and other ethnic groups residing in the United States and to explore the potential associations among these differences. Table 2 displays the result from the comparisons among eight subgroups in the Asian cohort: Filipino (19.1%), Japanese (26.9%), Chinese (23.9%), Hawaiian/Pacific Islander (6.9%), Korean (7.5%), Indian/Pakistani (3.0%), Vietnamese (5.6%), and others (7.1%). This study is one of the most comprehensive population-based analyses of CRC by ethnicity reported in the literature and looked at not only clinicopathologic factors (with 359 374 cases included) and also the lifestyle and screening data. Thus, it is reasonable to hypothesize that the relatively higher CRC survival for Indian/Pakistani patients results from better treatment. According to our study, overweight and obesity proportions in Asian subgroups were much lower than in NHW. Regarding aetiologic factors, high alcohol intake has been consistently associated with increased risk and moderate or high physical activity with decreased risk of CRC, and these factors vary by race, sex, and socioeconomic status (MurphyNone, 2011). Our results, which are based on a large, national population-based sample, show that there are differences in presenting clinicopathologic features between CRC patients of different races/ethnicities in the United States and that these differences may affect survival.
Gold summary	Colorectal cancer (CRC) diagnoses and disease-specific survival (DSS) vary between ethnic groups in the United States. However, few studies have assessed differences among Asian subgroups. The Surveillance, Epidemiology, and End Results (SEER) database was used to identify patients with invasive CRC between 1988 and 2008. Differences in clinicopathologic features and DSS rates were compared among Asian subgroups. The California Health Interview Survey was used to examine risk factors and screening patterns for CRC. The study included 359 374 patients with 8.4% Asian. Patients in all Asian subgroups were younger (median: 68 years) at diagnosis than non-Hispanic white (NHW) patients (median: 72 years). Most Asian subgroups, except Hawaiians, had better DSS than NHW patients although Asian subgroups had more advanced disease than NHW. Indian/Pakistani patients had a higher 5-year DSS than other Asian subgroups. Obesity proportions were lower in Asian subgroups (less than 50.2%) than in NHW (59.8%). Vietnamese men and Korean women had the lowest proportions of CRC screening. Advanced tumor stages were highly associated with worse DSS in each ethnicity group. High tumor grades were associated with worse DSS in NHW, Filipino, and Chinese. Older age at diagnosis was associated with worse DSS in most ethnicity groups except Hawaiian and Vietnamese. Disparities exist between Asians and NHW with CRC, and among various Asian subgroups. Differences in cancer clinicopathologic features, patients' behavioral habits, lifestyle, and screening patterns may explain some differences in CRC survival observed among ethnic groups.

Figure 2. A Comparison Between Our Model Summary and The Gold Summary In [23]

REFERENCES:

[1] Polatoğlu, Ilker, Tulay Oncu-Oner, Irem Dalman, and Senanur Ozdogan. "COVID-19 in early 2023: Structure, replication mechanism, variants of SARS-CoV-2, diagnostic tests, and vaccine & drug development studies." MedComm 4, no. 2 (2023): e228.

[2] Dehru, Virender, Pradeep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi, and Pawan Kartik. "Text summarization techniques and applications." In IOP Conference Series: Materials Science and Engineering, vol. 1099, no. 1, p. 012042. IOP Publishing, 2021.

[3] Ramos, Juan. "Using TF-IDF to determine word relevance in document queries." In Proceedings of the first instructional conference on machine learning, vol. 242, no. 1, pp. 29-48. 2003.

[4] Mihalcea, R., Tarau, P. "TextRank: bringing order into text". In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004).

[5] Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." Physical Review E 69, no. 6, pp.066-138, (2004).

[6] Satorra, A., Bentler, P.M. "A scaled difference chi-square test statistic for moment structure analysis". Psychometrika 66(4), 507–514 (2001).

[7] Wang, Kai, Jiahui Huang, Yuqi Liu, Bin Cao, and Jing Fan. "Combining feature selection methods with BERT: An in-depth experimental study of long text classification." In International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 567-582.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Springer, Cham, 2020.

[8] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In Proceedings of naacL-HLT, vol. 1, p. 2. 2019.

[9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30, pp. 5998–6008 (2017).

[10] Kieuvongngam, Virapat, Bowen Tan, and Yiming Niu. "Automatic text summarization of covid-19 medical research articles using BERT and GPT-2.", arXiv preprint arXiv:2006.01997, (2020) .

[11] Jabri, R. S., and E. Al-Thwaib. "Text Summarization Versus CHI for Feature Selection". Journal of Advances in Mathematics and Computer Science 22 (4), 1-8,(2017).

[12] Moradi, Milad, Georg Dorffner, and Matthias Samwald. "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization." Computer methods and programs in biomedicine 184, pp. 105-117, 2020.

[13] Moradi, Milad. "CIBS: A biomedical text summarizer using topic-based sentence clustering." Journal of biomedical informatics 88, pp. 53-61, (2018).

[14] Moradi, Milad, and Nasser Ghadiri. "Different approaches for identifying important concepts in probabilistic biomedical text summarization." Artificial intelligence in medicine 84, 101-116 (2018).

[15] Saggion, Horacio. "A robust and adaptable summarization tool." Traitement Automatique des Langues 49, no. 2 (2008): 68.

[16] TexLexAn: an open-source text summarizer. <http://texlexan.sourceforge.net/> (accessed 01/02/2019).

[17] Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, and Affandy Affandy. "Review of automatic text summarization techniques & methods." Journal of King Saud University-Computer and Information Sciences 34, no. 4 (2022): 1029-1046.

[18] Chiche, Alebachew, and Betselot Yitagesu. "Part of speech tagging: a systematic review of deep learning and machine learning approaches." Journal of Big Data 9, no. 1 (2022): 1-25.

[19] Lin, Chin-Yew. "Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?." In NII Testbeds and Community for Information Access Research (NTCIR). (2004).

[20] Lin, Chin-Yew. "ROUGE: A package for automatic evaluation of summaries." In-Text summarization branches out, pp. 74-81, (2004).

[21] Mitkov, Ruslan, ed. "The Oxford handbook of computational linguistics". Oxford University Press, 2005.

[22] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4, pp.1234-1240, (2020).

[23] Yi, M., Xu, J., Liu, P., Chang, G.J., Du, X.L., Hu, C.Y., Song, Y., He, J., Ren, Y., Wei, Y. and Yang, J. "Comparative analysis of lifestyle factors, screening test use, and clinicopathologic features in association with survival among Asian Americans with colorectal cancer." British journal of cancer 108, no. 7, pp.1508-1514 (2013).