**Proceedings of the ASME 2020
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2021
August 17-20, 2021, Virtual, Online**

IDETC2021-70866

ENGINEERING DOCUMENT SUMMARIZATION USING SENTENCE REPRESENTATIONS GENERATED BY BIDIRECTIONAL LANGUAGE MODEL

Yunjian Qiu
IMPACT Laboratory
Dept. of Aerospace & Mechanical Engineering
University of Southern California
Los Angeles, California 90089
yunjianq@usc.edu

Yan Jin*
IMPACT Laboratory
Dept. of Aerospace & Mechanical Engineering
University of Southern California
Los Angeles, California 90089
yjin@usc.edu
(*corresponding author)

ABSTRACT

In this study, the extractive summarization using sentence embeddings generated by the finetuned BERT (Bidirectional Encoder Representations from Transformers) models and the K-Means clustering method has been investigated. To show how the BERT model can capture the knowledge in specific domains like engineering design and what it can produce after being finetuned based on domain-specific datasets, several BERT models are trained, and the sentence embeddings extracted from the finetuned models are used to generate summaries of a set of papers. Different evaluation methods are then applied to measure the quality of summarization results. Both the automatic evaluation method like Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and the statistical evaluation method are used for the comparison study. The results indicate that the BERT model finetuned with a larger dataset can generate summaries with more domain terminologies than the pretrained BERT model. Moreover, the summaries generated by BERT models have more contents overlapping with original documents than those obtained through other popular non-BERT-based models. It can be concluded that the contextualized representations generated by BERT-based models can capture information in text and have better performance in applications like text summarizations after being trained by domain-specific datasets.

Keywords: Sentence embeddings, text summarizations, engineering context, knowledge capturing, language model, contextualized representations

1. INTRODUCTION

As the development of technology accelerates, large quantities of documents and papers are generated in almost all technical domains. As a result, it becomes challenging to efficiently capture the main knowledge and information from a vast amount of text documents. In recent years, there have been explorations to use automatic text processing techniques to process technical documents in the domains like medical, healthcare, and biology [1]. Automatic text summarization is a subfield of automatic text processing and natural language processing to deal with the problem of the overwhelming amount of text data [2]. Text summarization is a process of generating a summary that represents the most significant part of a document, such as a paper or multiple documents. In case the summarization is carried out for only one document, it is called *single-document* summarization, and for multiple documents, it is called *multi-document* summarization [3]. Furthermore, depending on how the summaries are constructed, there are *extractive* summarization and *abstractive* summarization [3, 4]. *Extractive* summarization is generated by existing sentences in the original text, and *abstractive* summarization is composed of new words and sentences which are different from those in the original documents seeking improved coherence of summary. Due to its relative simplicity, *extractive* summarization has often been applied to help identify the most important ideas in lengthy documents or papers.

Recently, researchers in the healthcare area, primarily the biomedical domain, have made great efforts on automatic text summarization in order to quickly grasp the main conclusions and findings in the paperwork like clinical reports without reading the whole text [5]. Initially, the research focused on sentence features like term frequency or position of sentences in the original text, and a number of techniques have been developed [4, 6]. Although these techniques are helpful to characterize the general relevance of the texts, they may not be sufficient to capture the most significant sentences in the text and generate a high-quality summary [5, 7]. Therefore, attempts have been made to extract domain knowledge from the original document, generate word presentations, and measure the similar information between the words [3, 7, 8]. These approaches have the potential to capture the semantic relationship and informative contents in sentences. As machine learning prospered in recent years, neural-network-based learning techniques have been applied to extract the domain knowledge through training based on large datasets [9, 10, 11, 12]. The neural-network-based approach allows the model to learn different features and map each word into vector representations in order to capture the semantic and syntactic meaning of the words and apply text understanding to text summarization [3, 10]. As an example, the contextualized word embeddings, which not only capture semantic meaning but also grasp contextualized meaning based on the surrounding context of words, have been widely applied to multiple downstream NLP tasks with desired performance [13, 14, 15]. Contextual word embeddings from the deep neural network language model have also been applied in text summarization, and the results are outstanding [16, 17, 18, 19].

In the engineering design area, the domain knowledge behind context is highly significant. It can be applied for design support as well as design ideation. In order to capture and reuse the domain knowledge, researchers often focus on information retrieval. Traditional keyword-based retrieval models can be used for literal matching but cannot meet the requirements on capturing the semantic information within the text [20, 21]. To address the challenge, researchers began to focus on knowledge retrieval like ontology-based retrieval to capture the ontological concepts and their relationships [22, 23, 24]. Although the ontology-based information retrieval approach can capture the semantic information to a certain extent, the "flat search" based approach is limited by its inability to "understand" the text in a high dimensional space where the words and sentences are cast together with "meaningful" relations. Partly due to the lack of language models and datasets, little work has been done on sentence-level knowledge capturing and its applications, such as text summarizations.

Inspired by the natural language processing (NLP) research and applications found in the biomedical domain, in this research, the contextual embeddings generated by language models are applied to capture the semantic meaning of the words and sentences in engineering documents and to generate text summarization of the documents. In this research, a language model called Bidirectional Encoder Representations from Transformers (BERT) is applied. BERT model is a Google-developed language model released in 2018 [15]. It uses attention mechanisms as well as a deep network architecture to learn and understand the unstructured text. After being pretrained using Masked Language Model and Next Sentence Prediction methods with over 100 million parameters, this model can capture the surrounding information of words and generate word representations that can dynamically change according to their positions [15]. BERT can also be finetuned to complete downstream tasks like question answering or sentence classification. Thanks to its unique and powerful architecture and extensive pretraining, the BERT model achieves the best performance over 11 NLP tasks. Based on this language model, one can use a much smaller dataset to finetune it to complete target-specific tasks. Contextual embedding can also be captured from different layers for tasks like text generation and text summarization [25, 26].

In this study, contextual embeddings for words and sentences are captured by the BERT models finetuned from given engineering design documents. Those representations are then applied to generate summarizations. In order to extract domain knowledge from unstructured text, three datasets with different sizes are created and labeled to finetune the BERT model. The outputs of contextualized language models are investigated by comparing them to the context-free methods. The comparison results demonstrate that contextual representations are able to capture domain knowledge after being finetuned with labeled data and can acquire important information from the original texts. The contributions of this research are:

- Demonstrated and evaluated the domain knowledge capture function of the BERT language model by collecting and creating domain-specific datasets to finetune the model and evaluating its effectiveness through comparative studies.
- Introduced a language-model-based approach, composed of contextual representations and clustering methods, to understand the text and select the most informative sentences for document summarization.
- Introduced summary evaluation methods and metrics from different angles and uncovered the underlying information behind the result of summarizations.

For simplicity, this study focuses on *extractive* summarization and investigates the applicability of the BERT models.

2. A SYSTEMATIC APPROACH TO DEALING WITH THE ISSUE OF TEXT SUMMARIZATION

In the engineering design area, it is valuable to identify and apply the useful information or rules underly past documents like design reports or papers. To acquire design knowledge from unstructured texts, researchers have focused more on word embeddings, or keyword search, for design creativity inspiration or rule generation. However, text understanding at the sentence level has rarely been used for knowledge acquisition due to the lack of benchmark datasets and adequate language models, poor training performance of models, and high computational burden. In addition, it is worth mentioning that text understanding is the principal step for researchers to extract domain knowledge and utilize the knowledge to process a large number of corpora. Therefore, there is a strong need for devising ways to train the language models to read and understand the unstructured texts and generate its "understanding" in a format that can quickly and effectively help human designers grasp the essential knowledge without having to read lengthy documents.

In this study, a systematic approach is proposed to investigate the language models capturing and learning specific domain knowledge from different datasets and generating corresponding summarizations. The results produced by the pretrained language model and finetuned language models are then analyzed and evaluated. Specifically, manually labeled datasets are created and used to finetune a BERT-based language model. During this process, the BERT model can learn to select significant sentences in the papers and capture the main idea underlying the sentences. In addition, using the K-Means method, the sentence embeddings extracted from the BERT model are clustered, and the sentences with the closest distance from the centroid of each cluster are selected and included in the final summarization. Since the BERT model can only deal with classification problems, the extractive summarization is considered as a binary classification problem where the labels, i.e., 1 and 0, are used to indicate whether a sentence should be included in the summary or not. Moreover, to generate text, sentence embeddings are extracted from the layers of neural networks in the BERT model, and then the K-Means method is used to create different clustering which is formed by those sentence embeddings. Here the number of clusters represents the number of sentences included in the summary.

In this way, the BERT models can capture the domain knowledge during the finetuning process, and sentence embeddings extracted from BERT models will contain those informative content. After using the clustering method, corresponding summarization can finally be generated. In order to compare the performance of language models with and without being trained, the same procedures after finetuning are applied to pretrained BERT model as well. The flow of the information about this systematic approach is shown in Figure 1. below.

2.1 Data collection and preprocessing

A desired finetuned BERT model should be able to select critical sentences in one paper and generate corresponding sentence embeddings used for summarization. For attaining such a model, the first step is to create a dataset that can be used to train and test the BERT model to capture the main idea in the text. Due to the lack of benchmark datasets in the engineering design area, one sample dataset is created manually to show how the BERT model learns the engineering-specific knowledge and how altering parameters impact summarization results.

The sentences in the raw dataset are collected from papers about additive manufacturing which is a subfield in engineering. In order to assess the influence of the size of datasets, three datasets with distinct sizes are created. Correspondingly, 38, 60, and 172 most recent papers are selected from ScienceDirect and are considered as original data. To train the BERT model to learn from the sentences that can represent the informative content in papers and automatically generating summarization, only the main parts of the selected papers are captured from original documents, including *abstract*, *introduction,* and *conclusion* sections. Due to the requirements of finetuning the BERT model, the paragraphs need to be tokenized up into individual sentences using the NLTK toolkit [27]. Finally, there are 505, 2020, and 6167 sentences in the three raw datasets, respectively.

In order to reduce the noise in the datasets, the reference symbols, the caption of figures and tables as well as mathematical equations are removed. Moreover, since the NLTK toolkit tokenizes the sentences out of paragraphs by period, it may result in incorrect splitting work when it comes to situations like ‘Fig.’ or ‘etc.’. Therefore, efforts are also made to combine separated sentences due to incorrect splitting. Besides, some authors summarized their main ideas or findings in the table format, which cannot be directly captured by the NLTK toolkit. Under that circumstances, the main ideas or findings inside the tables should be extracted by manual work. In addition, some irrelevant sentences like figure captions are removed.

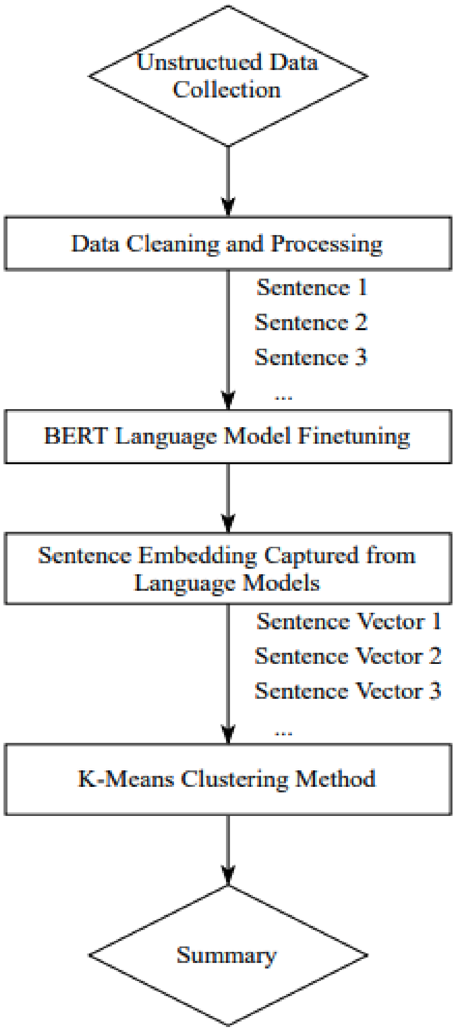
Unstructued Data Collection
Data Cleaning and Processing
Sentence 1
Sentence 2
Sentence 3
...
BERT Language Model Finetuning
Sentence Embedding Captured from Language Models
Sentence Vector 1
Sentence Vector 2
Sentence Vector 3
...
K-Means Clustering Method
Summary

Figure 1: *The process of text summarization generation using BERT language models.*

For training the BERT model to capture the domain knowledge and automatically generate extractive summarization, the sentences tokenized from original content need to be labeled. In this study, extractive summarization is defined as a classification case. For each sentence, it is label as {1,0} to indicate whether the sentence should be included in the summarization. Sentences containing the most important information are labeled as 1. For instance, sentences in abstract and conclusion parts which display the main ideas and findings of paper would be labeled as 1, while sentences in the introduction part which convey information about related background would be considered as less important sentences and labeled as 0. Additionally, to meet the requirements of the BERT model, for each sentence, the tokens [CLS] and [SEP] are inserted at the start and end of the sentences correspondingly. Finally, in order to maintain the BERT model learning the most important information, the standards for selecting important sentences are rigid. The proportion of binary labels {1, 0} is around 1:2.

2.2 Language model

BERT [15] is a language representation model developed by Google. This new model is different from past language models like RNNs and outperforms other language models in over 11 NLP tasks. Since it can be finetuned to complete specific tasks with a relatively small dataset and can map words and sentences to contextualized representations, it is chosen as the language model in this study to process the unstructured text.

The reason why BERT can have the best performance in NLP tasks is because of its model structure and input/output representations. Firstly, the model structure of BERT is distinct from other models in respect of its robustness. The main part which guarantees contextual learning is the transformer, which is an attention mechanism. It can convert text to word vectors that then are processed in the neural network. For maintaining bidirectional learning, the inputs first flow through an attention layer which guarantees the encoder learning from the surrounding context. Then the output of the attention layer will be processed to a neural network. Since the attention mechanism contains multi attention heads, it can significantly enhance the computational performance and increase training accuracy even with small datasets. As for the output layer, it will process the sequence of words and convert them into vectors. Like the input process, the attention layer will help the decoder concentrate on the position of the input sequence. After the vectors are output from the decoding layer, they would be processed by the final linear and SoftMax layer and turn the vectors into words. Therefore, the input representation can be constructed by three parts to better capture the position as well as the contextual meaning of the input sequence. Token embeddings, segment embeddings, and position embeddings will be summed and considered as the final input representations. Then they can be utilized to complete downstream language tasks like sentence classification. Moreover, in order to map the text to contextualized representations, during pretraining process, masked language models and next sentence prediction are used to capture the surrounding information of the words. In this study, the BERT model is used for converting words and sentences into contextualized vectors and completing the summarization tasks.

Currently, there are several different pretrained BERT models which are trained under distinct layers. Generally, pretrained BERT model can be categorized as a BERT-base model, which has 12 layers, and the BERT-large model, which has 24 layers. In this paper, the BERT-base model is applied because of computational efficiency.

2.3 Experimental method

In this study, the BERT model and K-Means method are combined to realize generating summarization automatically. Since the BERT model cannot be used for text generation directly, it is used to generate sentence embeddings. The sentence embeddings, represented as vectors, are considered as input of a K-Means method. Using the K-Means method with sentence embeddings can generate several clusters, and the sentences that are nearest to the centroids of clusters will be selected and be included in the final summarization.

In order to evaluate the influence of the size of the dataset on the BERT training results, three different datasets are created with 505, 2020, and 6167 sentences, respectively. The three datasets are used to train three distinct BERT models. In these training datasets, validation data are selected to help monitor the entire training process. Generally, the proportion of the training dataset and validation dataset is 9:1. Besides, to explicitly show the accuracy of those training models, the same testing dataset is applied. The testing dataset contains 102 sentences. During the finetuning process, hyperparameters need to be set to help the BERT model achieve the best performance. The authors of the BERT model recommend using 2-4 epochs to train the BERT model [15]. In this study, 4 epochs are chosen to finetune the BERT model. Moreover, the recommending learning rate is from 2e-5 to 5e-5; after several experiments, it becomes clear that learning rate 2e-5 leads to the highest accuracy using a specific dataset. Moreover, considering the size of the dataset, the batch size is set to 16. Batch size 8 is small, and batch size 32 runs out of memory limitations. Table 1 below displays the information about parameters of BERT models during the finetuning process.

Table 1: *Parameters of BERT model in finetuning process*

	Learning Rate	Batch Size	Numbers of Epochs	Layers
BERT model	2e-5	16	4	12

After obtaining the finetuned BERT model, sentence embedding can be captured from the network architecture. In this study, word representations are extracted in the last two layers of the neural network, and sentence embeddings are generated by averaging the word representations to convert the different lengths of sentences into fixed-length vectors. In the BERT model, sentence embeddings are N×E vectors where N represents the number of sentences and E represents the dimension of embeddings. Usually, the default embedding dimension is 768.

For dealing with different sentence embeddings and capturing the sentences which can represent the main idea, K-Means method [28] is applied to generate clusters. Sentences with similar information will be collected into one cluster based on their sentence embeddings. After the clustering of sentences being generated, the centroid of clustering will be calculated, and the sentence with the closest Euclidean distance from the centroid will be chosen as the main sentence embeddings. Finally, all the sentence embeddings are combined, and their corresponding sentences will be included in the final summarization. In this study, in order to avoid poor clustering results, k-means++ is chosen to set up initialization. Moreover, the influence of dimensions of sentence embeddings on clustering results will be investigated.

2.4 Evaluation

After the summarizations are generated by BERT-based methods, the results will be compared with summarizations created from non-BERT-based methods. Different evaluation methods will be applied to measure the quality of the summaries.

In this paper, the performance of the BERT-based approach is compared with the three most popular non-BERT-based summarizers, i.e., KL-Sum algorithm, TextRank, and Latent Semantic Analysis (LSA). KL-Sum algorithm [29] stands for Kullback-Lieber Sum algorithm and is a content-based approach that selects a sequence of sentences from text based on unigram distribution. The concept of KL divergence is applied to measure the difference of probability distribution of distinct contexts in order to discover their similarity. TextRank [30] is a graph-based algorithm and is an unsupervised approach. It ranks sentences on the basis of their cosine similarity scores and extracts top sentences for summarization. LSA [31] is a topic-based approach that evaluates the significance of sentences by their singular value decomposition (SVD) values. Random baseline, which selects sentences in original text randomly, will also be compared as a benchmark.

The evaluation of the generated summaries is an unsolved task for the research community and is still being discussed. While there are still many problems concerning the methods and types of evaluations, two evaluation methods are chosen for evaluating the performance of summarization systems in this paper. Mainly, the evaluation systems can be classified into two categories. One is called *extrinsic evaluation,* where the quality of summary is evaluated based on the influence brought by the summary. Another is called *intrinsic evaluation,* where the quality of summary is directly assessed based on the content of summarization like word frequency or longest matching sequences [32]. For instance, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [33] is a widely used intrinsic evaluation due to its efficacy. In ROUGE, *precision*, *recall,* and *F-score* would be applied as an evaluation metric to evaluate the quality of the summary. It generates this evaluation metric by comparing the standard summary or reference summary and automatically generated summary. Based on different criteria, it measures the overlapping information between reference summary and generated summary. Higher scores mean that more overlapping content is captured. Specifically, the recall score refers to the proportion of overlapping content presented in the reference summary; and the precision score refers to the proportion of overlapping content presented in generated summary. In this experiment, ROUGE-1, ROUGE-2, and ROUGE-L scores are utilized to assess the summary quality since these scores can work well in a single-document summary [34].

One disadvantage of using ROUGE, however, is that the standard summary is always required to compare with the generated summary. It would be difficult to find an ideal summary since there are no formal rules to establish one [35]. Commonly, researchers may use a human-made summary or abstract of papers as standard. Despite that, it may be biased to merely measure the overlapping content between the standard summary and the generated summary since the authors may avoid using the same expressions in the main content, which can decrease the possibility of overlapping. They tend to utilize synonyms or change expressions to describe the same idea in order to maintain word diversity and legibility. Therefore, another evaluation method for statistical-based evaluation, which only focuses on generated summary, can be applied in this experiment as a supportive approach. According to [36, 37], keywords in a document can represent the most significant idea of its content, meaning that a summarization would contain more high-frequent words. Consequently, other than ROUGE, word frequency measurement, which is a statistical-based method, would also be considered as an evaluation method to measure the quality of the summary. Specifically, after removing stop words, sentences with more most frequent words will be assigned higher scores, and the average of those sentence scores will be the final score of the generated summary. In order to avoid the potential issue brought from long-length sentences, the score of each sentence will be divided by the number of words in the sentences.

3. RESULTS

3.1 Finetuning results

When the size of the training dataset is different, the finetuned BERT model can show different results. In order to assess the influence of the size of training datasets, two additional datasets with 1000 sentences and 1500 sentences are created for the experiment. And a small testing dataset that contains 102 sentences is also applied to measure the testing accuracy of models. Table 2 shows the information about the training accuracy and testing accuracy for different BERT models. For the BERT model finetuned by 505 sentences, as Table 2 shows below, the training accuracy of it is about 69% which is highly increased comparing to pretrained BERT model. As the size of the dataset increases, the accuracy of BERT models is also improving. The comparison results indicate that the size of the training dataset is one vital factor that influences the performance of BERT models. As the size of the dataset increases, the speed of accuracy improvement becomes slower. Moreover, from Figures 2 to 5., which present the change of loss and accuracy during the training process, it can be seen that during the training process, when the dataset contains only 505 sentences, as shown in Figure 2, there is only a slight change of the accuracy over training epochs. But when the dataset is enlarged by 500 more sentences shown in Figure 3, the trend of accuracy improvement becomes a linear pattern with respect to the training epoch, indicating the effect of learning of the BERT model.

Table 2: *Training & testing accuracy of different BERT models*

	Pretrained model	Finetuned with various # of sentences				
		505	1000	1500	2012	6167
Training	X	0.691	0.763	0.808	0.832	0.858
Testing	0.3	0.554	0.708	0.722	0.722	0.775

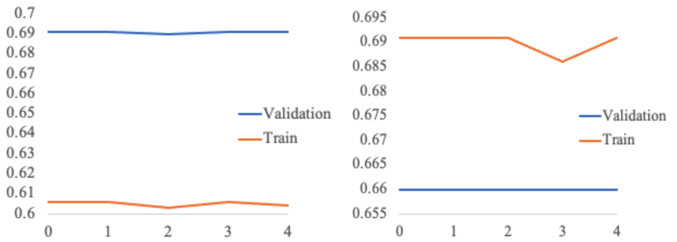

Figure 2: *Training loss (left) and accuracy (right) of BERT model trained by 505 sentences*

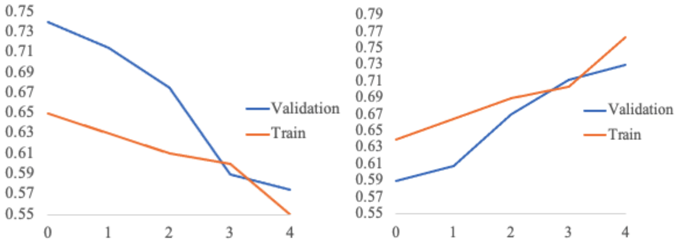

Figure 3: *Training loss (left) and accuracy (right) of BERT model trained by 1000 sentences*

Under the circumstance where the size of the dataset keeps expanding by 500 sentences, the accuracy of the finetuned model gains significant improvement. However, when the size of the dataset reaches 2000 sentences and beyond, the accuracy of the finetuned model did not increase as significantly as before. This result indicates that when the size of the dataset has reached a large enough level, i.e., around 2000 in this study, further increasing the size with 500-sentences increment can only result in moderate accuracy gains. Moreover, as the size of the dataset increased to 5000 sentences and beyond, the accuracy improvement almost stalled. Nevertheless, as Table 1 shows, the finetuned BERT model with 6167 sentences achieved the best performance in the testing dataset, demonstrating that when the dataset is large, the model can extract and learn more knowledge from the sentences.

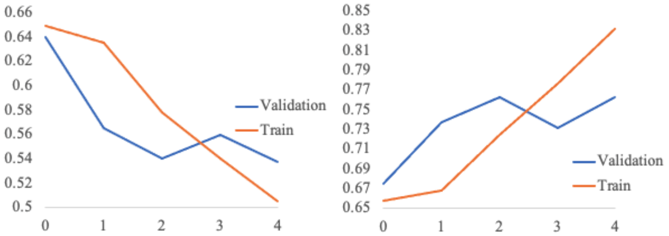

Figure 4*: Training loss (left) and accuracy (right) of BERT model trained by 2012 sentences*

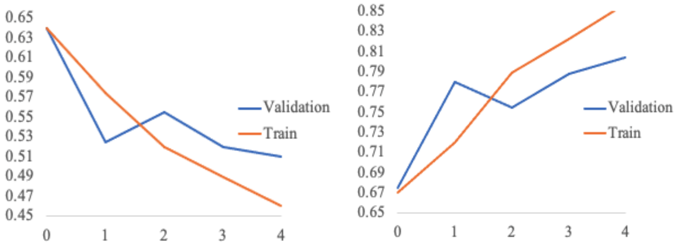

Figure 5*: Training loss (left) and accuracy (right) of BERT model trained by 6167 sentences*

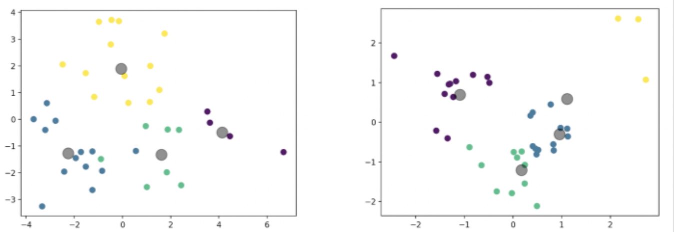

Figure 6: *Representation of 2D sentence embeddings captured from pretrained model (left) and finetuned model by 500 sentences (right)*

Before sentence embeddings are extracted from BERT models, visualizations of sentence embeddings under 2D are generated to show the differences between the pretrained BERT model and the finetuned BERT models. Figure 6 illustrates the visualization of sentence embeddings in a 2D coordinate. Based on the plot, it can be seen that the sentence embedding has significantly changed after the model is finetuned. Therefore, a hypothesis can be made that the summarizations generated by the finetuned models with higher accuracy possess more important information and terminologies in the additive manufacturing domain compared to pretrained model. More detailed experiments and evaluations will be discussed in the next section.

3.2 Summarization evaluation results

According to what Lin [33] demonstrated, the critical number of documents for single-document summarization evaluation is 86. Therefore, for our testing dataset, 101 papers that were published in recent years are randomly selected from ScienceDirect. These papers have the same focus on additive manufacturing as the papers in the training datasets. Only *abstract*, *introduction*, and *conclusions* are selected from the original papers for capturing the most important contents. Among them, the sentences in the *introduction* and *conclusions* sections are extracted for summarization using different algorithms, while the *abstracts* of papers are applied for standard summarization, which is then compared with the generated summary in ROUGE evaluation. In the statistical analysis evaluation, only the generated summaries are evaluated by measuring their word occurrence. Moreover, during the evaluation process, the number of sentences in the generated summaries is maintained the same as that in the reference summary for meaningful comparison.

In order to compare the performance of different models, the same testing dataset is applied to the BERT-based model and

Table 3: *Summaries generated by sentence embeddings under different dimensions*

	Pretrained Model
2D	Unfortunately, as-printed surfaces originating from AM are rough and incapable of functioning as mating surfaces in a product assembly. Herein, AM can naturally produce a high density of volumetric-porosity defects and unique microstructure characteristics, e.g. preferred crystallographic textures, and, gradients in grain size. Addressing this knowledge gap requires an in-depth understanding of the mechanics of finishing in surface texture/microstructure /defect combinations that originate from AM. These insights are subsequently used to create a framework whose utility in optimizing finishing processes are discussed.
4D	Finishing of components originating from additive manufacturing (AM) is critically important for providing them with adequate tolerances and fatigue life. Using these insights, a finite element based numerical framework of surface deformation of additively manufactured IN718 is created. An attempt is made to delineate effects arising from surface roughness, microstructure gradients, and porosity defects. These insights are subsequently used to create a framework whose utility in optimizing finishing processes are discussed.
20D	Using these insights, a finite element based numerical framework of surface deformation of additively manufactured IN718 is created. These processes are used to create surfaces with tighter geometric control, reduced roughness, or residual compressive stresses. An attempt is made to delineate effects arising from surface roughness, microstructure gradients, and porosity defects. These insights are subsequently used to create a framework whose utility in optimizing finishing processes are discussed.
768D	Optimization of finishing processes is however challenging for AM components as their mechanics of deformation are complicated by microstructure/defect/ roughness combinations present in as-received surfaces. In this work, the mechanics of surface deformation in additively manufactured IN718 is studied via indentation. Hence, the surfaces of AM parts are typically subject to primary machining processes, peening processes, or secondary machining processes that use loose abrasives. An attempt is made to delineate effects arising from surface roughness, microstructure gradients, and porosity defects.

other non-BERT-based approaches. 101 scores for corresponding papers are averaged as the final scores for the summarizing approaches. The scores of those summaries from distinct evaluation methods are listed below in Table 5, which can represent the differences in the performance of these summarization methods.

3.3 Automatic evaluation

The ROUGE evaluation is used for two purposes in this paper. First, it is used to assess the influence of the size of dimensions used for sentence embeddings, and secondly, it is applied to compare ROUGE scores of summaries generated by pretrained BERT model and finetuned BERT models.

3.3.1 Parameterization

In order to identify proper parameter settings for achieving the best performance, the impacts of different sizes of dimensions for sentence embeddings are investigated. As shown in Table 3, summaries generated by sentence embeddings with different dimension sizes are distinct, indicating that dimensionality reduction that can be applied in summarization to increase computational efficiency may influence the final output.

Generally, researchers choose 2D sentence embedding to complete summarizing work for data visualization and computational efficiency [4]. However, the extent of loss of performance after reducing dimensions of sentence embeddings still need to be investigated. In this experiment, to evaluate the quality of summaries under different dimension sizes, ROUGE evaluation is applied to measure the impacts of dimensionality reduction. Moreover, only 2D, 4D and 20D sentence embeddings are selected to compare with original 768D sentence embeddings given that the dimensionality reduction algorithm requires the component setting be no larger than the number of samples. Since the minimum number of samples in the dataset is 22, the components are set less than 22 dimensions. Table 4. below shows means scores of summaries with different dimensions as well as the underlying language models. According to the results shown in Table 4, the loss of performance exists during dimensionality reduction. Specifically, sentence embeddings

Table 4: *Mean value under different dimensions*

Dimensions	**Pretrained BERT**			**Finetuned BERT (500)**			**Finetuned BERT (2021)**			**Finetuned BERT (6067)**		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
2D	0.394	0.098	0.285	0.403	0.112	0.305	0.343	0.094	0.265	0.389	0.096	0.294
4D	0.414	0.119	0.308	0.342	0.078	0.250	0.352	0.079	0.267	0.355	0.083	0.259
20D	0.400	0.107	0.297	0.365	0.105	0.292	0.370	0.077	0.272	0.387	0.106	0.290
768D	**0.421**	**0.137**	**0.319**	**0.411**	**0.129**	**0.308**	**0.423**	**0.135**	**0.323**	**0.427**	**0.144**	**0.323**

* the number in brackets represents the size of dataset.

Table 5: *Mean value of ROUGE-1, ROUGE-2 and ROUGE-L scores of BERT-based summarizers and non-BERT-based summarizers*

	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
Summarizers	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
Pretrained-BERT	0.432	0.420	0.421	0.137	0.136	0.134	0.335	0.304	0.317
Finetuned BERT (500)	0.387	0.448	0.411	0.121	0.141	0.129	0.300	0.320	0.308
Finetuned BERT (2012)	0.394	0.457	0.423	0.126	0.146	0.135	0.314	0.332	0.323
Finetuned BERT (6167)	0.405	0.452	**0.427**	0.135	0.154	**0.144**	0.316	0.328	**0.323**
Text Rank	0.502	0.359	0.415	0.168	0.118	0.138	0.361	0.278	0.311
KL-Sum	0.370	0.427	0.392	0.114	0.135	0.122	0.289	0.337	0.308
LSA	0.375	0.382	0.378	0.108	0.109	0.108	0.317	0.284	0.298
Random Baseline	0.380	0.369	0.374	0.109	0.104	0.105	0.303	0.272	0.284

* the number in brackets represents the size of the dataset.

with 768D can generate summaries with higher scores, meaning that the 768-dimension sentence embedding model captures more information and has a better performance comparing to other dimensional models. In other words, high-dimensional models can acquire the most representative information of the papers. Therefore, in this experiment, all the summaries are generated by 768D sentence embeddings in order to ensure the best results.

3.3.2 Comparisons among different summarizers

Table 5 presents the mean value of ROUGE-1, ROUGE-2 and ROUGE-L scores acquired from BERT-based summarizers and other non-BERT-based approaches. As Table 5 shows, the performance of the finetuned BERT model by 6167 sentences exceeds other BERT-based summarizers and non-BERT-based summarizers with regard to ROUGE-1, ROUGE-2, and ROUGE-L scores.

Taking ROUGE-2 score as an example, the scores of the summaries generated from the finetuned BERT model by 6167 sentences are over 11.6% higher than those generated from other BERT-based models. Moreover, comparing to other non-BERT-based summarizers, the mean scores obtained from finetuned BERT model by 6167 sentences are about 37.1% higher than non-BERT-based approaches.

More specifically, when comparing the finetuned BERT models with other summarizers in terms of *precision* score, the result shows that the finetuned BERT models with different dataset sizes almost all outperform other text summarizers, while the *recall* scores of finetuned BERT models are relatively low or indifferent. In addition, considering the comparisons among three finetuned BERT models with different dataset sizes, the mean scores increased as the size of datasets was enlarged, which is also consistent with the corresponding accuracy of BERT training models discussed above. For instance, when the size of the dataset was enhanced from 500 sentences to 6167 sentences, the mean value of ROUGE-L scores rose about 4.9%. Also, the finetuned BERT model with the largest dataset presented the best performance comparing to pretrained BERT model.

3.4 Statistical evaluation

Table 6 below presents the evaluation results obtained from different BERT-based models in terms of *word frequency* measurements. As shown in the data, the mean scores have improved from the pretrained BERT model to the finetuned BERT models. In comparisons among three different finetuned BERT models, as the size of the dataset increases, the performance of finetuned BERT models is enhanced. Specifically, the mean scores of summarizers with a larger dataset can be 14.50% higher than others.

Table 6: *Average word frequency score of Summarization Generated by Different BERT Models*

* the number in brackets represents the size of the dataset.

Summarizers	Mean	Min	Max
Pretrained BERT	3.432	2.475	4.660
Finetuned BERT (500)	3.785	2.175	5.572
Finetuned BERT (2012)	3.790	2.414	5.817
Finetuned BERT (6167)	3.930	2.524	5.579

4. DISCUSSION

Summarizations based on dimensions of contextualized representation. Different settings of the parameters and the size of dimensions of sentence embeddings can lead to different results. In this study, 2D, 4D, 20D, and 768D sentence embeddings are evaluated. Based on the ROUGE results, it has been shown that sentence embeddings under 768D can generate summaries with better performance than others. This demonstrates that the sentence embeddings with higher dimensions can capture more information of the text documents.

Domain knowledge capturing by model finetuning. Comparing pretrained BERT models and the finetuned BERT models, the results illustrated that the BERT model can increase the number of keywords in summaries after finetuned by datasets containing specific domain knowledge. Moreover, based on the results from ROUGE evaluation, it can be seen that the summaries generated by the BERT model finetuned with the largest dataset have a greater overlapping extent than those generated by other BERT-based models and non-BERT-based models. This result demonstrates that the word representations in the finetuned BERT models can capture the informative context even better than the pretrained BERT model thanks to the additional and domain-specific training. Besides, in comparing recall scores and precisions scores of BERT-based summarizers, the result shows that the recall scores decreased while the precisions scores were highly enhanced after the finetuning process. This situation happens because the finetuned BERT models are trained to learn how to select the most significant sentences in the document rather than to learn specifically for word representations like the pretrained BERT model. Due to the rigid selection of important sentences in training datasets, the finetuned BERT models can be critical when choosing the most informative sentences. This has made the finetuned BERT models have high precision scores and low recall scores. The high precision score means the summary generated by finetuned BERT models has more overlapping words with the original abstract comparing to other models. Meanwhile, the low recall score illustrates that much information in the original text is insignificant from the perspective of the finetuned BERT models. Therefore, the summaries do not capture relevant but insignificant sentences. That explains why the summaries generated by the finetuned BERT models have more keywords and higher precision scores but lower recall scores comparing to the pretrained BERT model.

Performance of sentence representations from various sizes of dataset. Comparing the finetuned BERT models for different finetuned BERT models with distinct dataset sizes, the model trained by larger datasets outperforms those by smaller ones. According to ROUGE scores and statistical scores, the finetuned BERT models with the largest dataset can always achieve the best performance. This indicates that being trained with larger datasets, the finetuned BERT model can have higher accuracy in selecting the most important sentences and capturing more domain knowledge from the original texts.

Significant information contained in contextualized representations. By comparing BERT-based models and non-BERT-based models, it can be seen that after the words and sentences being mapped to contextual representations, the performance of BERT-based models can exceed most non-BERT-based models such as TextRank and latent semantic analysis (LSA), especially in terms of ROUGE-1 score. Moreover, the higher precision scores of finetuned BERT-based models indicate that they are able to capture more significant sentences in the original texts and the summaries generated by BERT-based models are highly informative.

5. CONCLUSIONS AND FUTURE WORK

In this study, sentence embeddings that convert unstructured text to multidimensional vectors are extracted from BERT models and are then used in K-Means method to capture the main idea of different papers and generate summaries. The evaluation results of the BERT models, together with other non-BERT models, indicate that the BERT model can increase the number of keywords in summaries after finetuned by datasets containing specific domain knowledge. Moreover, compared to the original abstract in papers, the summaries generated by finetuned BERT models have a greater overlapping extent than those generated by pretrained BERT models and other non-BERT-based models, which demonstrates that the word representations in finetuned BERT models can capture the informative context effectively.

Based on the results and discussions described above, it can be concluded that contextual embeddings can enhance the performance in Natural Language Processing (NLP) tasks like text summarization. In addition, finetuning process can increase the ability of BERT models to capture domain knowledge and apply the knowledge in word and sentence representations. Those contextual representations can capture semantic and contextual information and have great potential for processing other NLP tasks in different domains. From an engineering support point of view, the high effectiveness of finetuned BERT models has opened ways to developing extensive NLP tools to support engineering knowledge capture, personal NLP-based design assistance, and engineering collaboration.

In this paper, the contextualized embeddings are only used for document summarization through domain knowledge capture-based finetuning. Future work includes introducing human expert-based summary evaluation, exploring the features of the sentence embeddings by examining the clustering properties, and going beyond summarization by identifying design activities and thinking processes as NLP application tasks for the domain-specific finetuned BERT models. The long-term goal is to realize highly "intimate" computer-aided design by using BERT models to augment engineers' working and thinking processes.

REFERENCES

[1] Fleuren, W. and Alkema, W., 2015, "Application of text mining in the biomedical domain," *Methods*, 74, pp.97-106.

[2] Ferreira, R., de Souza Cabral, L., Lins, R., Pereira e Silva, G., Freitas, F., Cavalcanti, G., Lima, R., Simske, S. and Favaro, L., 2013, "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, 40(14), pp.5755-5764.

[3] Lloret, E. and Palomar, M., 2012, "Text summarisation in progress: a literature review," *Artificial Intelligence Review*, *37*(1), pp.1-41.

[4] Mishra, R., Bian, J., Fiszman, M., Weir, C.R., Jonnalagadda, S., Mostafa, J., and Del Fiol, G., 2014, "Text summarization in the biomedical domain: a systematic review of recent research," *Journal of biomedical informatics*, *52*, pp.457-467.

[5] Reeve, L. H., Han, H., & Brooks, A. D., 2007, "The use of domain-specific concepts in biomedical text summary-zation," *Information Processing & Management*, *43*(6), 1765-1776.

[6] Yao, J. G., Wan, X., & Xiao, J., 2017, “Recent advances in document summarization,” *Knowledge and Information Systems*, *53*(2), 297-336.

[7] Plaza, L., Díaz, A., & Gervás, P., 2011, “A semantic graph-based approach to biomedical summarization,” *Artificial intelligence in medicine*, *53*(1), 1-14.

[8] Ji, X., Ritter, A., & Yen, P. Y., 2017, “Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews,” *Journal of biomedical informatics*, *69*, 33-42.

[9] Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D., 2014, “Extractive summarization using continuous vector space models,” In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pp. 31-39.

[10] Camacho-Collados, J., & Pilehvar, M. T., 2018, “From word to sense embeddings: A survey on vector representations of meaning,” *Journal of Artificial Intelligence Research*, *63*, 743-788.

[11] Cheng, J., & Lapata, M., 2016, “Neural summarization by extracting sentences and words,” *arXiv preprint arXiv:1603. 07252*.

[12] Alami, N., Meknassi, M., & En-nahnahi, N., 2019, “Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning,” *Expert systems with applications*, *123*, 195-211.

[13] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L., 2018, “Deep contextualized word representations,” *arXiv preprint arXiv:1802.05365*.

[14] Akbik, A., Blythe, D., & Vollgraf, R., 2018, “Contextual string embeddings for sequence labeling,” In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638-1649.

[15] Devlin, J. et al., 2019, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” *NAACL-HLT*.

[16] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J., 2020, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, *36*(4), 1234-1240.

[17] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M., 2019, “Publicly available clinical BERT embeddings,” *arXiv preprint arXiv:1904. 03323*.

[18] Si, Y., Wang, J., Xu, H., & Roberts, K., 2019, “Enhancing clinical concept extraction with contextual embeddings,” *Journal of the American Medical Informatics Association*, *26*(11), 1297-1304.

[19] Peng, Y., Yan, S., & Lu, Z., 2019, “Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets,” *arXiv preprint arXiv:1906.05474*.

[20] Beigbeder, M., & Mercier, A., 2005, “An information retrieval model using the fuzzy proximity degree of term occurences,” In *Proceedings of the 2005 ACM symposium on Applied computing*, pp. 1018-1022.

[21] Castells, P., Fernandez, M., & Vallet, D., 2006, “An adaptation of the vector-space model for ontology-based information retrieval,” *IEEE transactions on knowledge and data engineering*, *19*(2), 261-272.

[22] Zhang, X., Hou, X., Chen, X., & Zhuang, T., 2013, “Ontology-based semantic retrieval for engineering domain knowledge,” *Neurocomputing*, *116*, 382-391.

[23] Sanya, I. O., & Shehab, E. M., 2015, “A framework for developing engineering design ontologies within the aerospace industry,” *International Journal of Production Research*, *53*(8), 2383-2409.

[24] Zhang, C., Zhou, G., Lu, Q., & Chang, F., 2017, “Graph-based knowledge reuse for supporting knowledge-driven decision-making in new product development,” *International journal of production research*, *55*(23), 7187-7203.

[25] Zhang, Haoyu, Jianjun Xu, and Ji Wang., 2019, “Pretraining-based natural language generation for text summarization,” *arXiv preprint arXiv:1902.09243*.

[26] Miller, Derek, 2019, “Leveraging BERT for Extractive Text Summarization on Lectures,” *ArXiv* abs/1906.04165: n. pag.

[27] Loper, E., & Bird, S., 2002, “NLTK: the natural language toolkit,” *arXiv preprint cs/0205028*.

[28] Bradley, P. S., & Fayyad, U. M., 1998, “Refining initial points for k-means clustering,” In *ICML*, Vol. 98, pp. 91-99.

[29] Haghighi, Aria, and Lucy Vanderwende., 2009, “Exploring content models for multi-document summarization,” *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[30] Mihalcea, R., & Tarau, P., 2004, “Textrank: Bringing order into text,” In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404-411.

[31] Ozsoy, Makbule & Alpaslan, Ferda & Cicekli, Ilyas., 2011, “Text summarization using Latent Semantic Analysis,” *J. Information Science,* 37. 405-417. 10.1177 /0165551511 408848.

[32] Inderjeet, M. A. N. I., 2009, “Summarization evaluation: an overview,” In *Proceedings of the NTCIR Workshop*, Vol. 2.

[33] Lin, Chin-Yew., 2004, “Rouge: A package for automatic evaluation of summaries,” *Text summarization branches out*.

[34] Lin, C. Y., 2004, “Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?” In *NTCIR*.

[35] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K., 2017, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*.

[36] Brysbaert, Marc & Mandera, Paweł & Keuleers, Emmanuel., 2018, "The word frequency effect in word processing: A review update," *Current Directions in Psychological Science*. 27. 10.1177/0963721417727521.

[37] Gupta, V., & Lehal, G. S., 2010, "A survey of text summarization extractive techniques," *Journal of emerging technologies in web intelligence*, *2*(3), 258-268.