Engineering Document Summarization: A Bidirectional Language Model-Based Approach

Yunjian Qiu
Department of Aerospace and Mechanical Engineering,
University of Southern California,
3650 McClintock Avenue, OHE 400,
Los Angeles, CA, 90089-1453
e-mail: yunjianq@usc.edu

Yan Jin[1]
Department of Aerospace and Mechanical Engineering,
University of Southern California,
3650 McClintock Avenue, OHE 400,
Los Angeles, CA, 90089-1453
e-mail: yjin@usc.edu

In this study, the extractive summarization using sentence embeddings generated by the finetuned Bidirectional Encoder Representations from Transformers (BERT) 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 data sets, 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 machine evaluation method Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and a human-based 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. The experimental results indicate that the BERT-based method can provide better and more informative summaries to engineers. It has also been demonstrated 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 data sets. [DOI: 10.1115/1.4054203]

Keywords: sentence embeddings, text summarizations, engineering context, knowledge capturing, language model, contextualized representations, artificial intelligence, computer-aided design, data-driven engineering, engineering informatics

1 Introduction

As the development of technologies 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 refers to generating a summary that represents the most significant part of a document, such as a paper or multiple documents. Depending on how the summaries are constructed, there are extractive summarization and abstractive summarization [3,4]. Extractive summarization generates summaries by using existing sentences in the original texts, whereas abstractive summarization composes summaries with new words and sentences that are different from those in the original documents for improved coherence. Due to its relative simplicity, extractive summarization has often been applied to help identify the most important ideas in lengthy documents or papers.

Inspired by the natural language processing (NLP) research and applications found in the biomedical domain [5–11], 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. More specifically, a language model called Bidirectional Encoder Representations from Transformers (BERT) is applied. The BERT model uses attention mechanisms as well as a deep network architecture to capture the surrounding information of words and generate word representations that can dynamically change according to their positions [12]. From an engineering design support perspective, one important feature of BERT is that it can be finetuned to complete downstream tasks like question-answering or sentence classification, making it possible to carry out domain-specific tasks. Thanks to its unique and powerful architecture and extensive pretraining, one can use a much smaller data set to finetune it to complete target-specific tasks, which can avoid manually generating extremely large labeled data sets needed for composing one's own NLP model. Contextual embedding can also be captured from different layers for tasks like text generation and text summarization [13,14].

The BERT model has been pretrained based on a vast range of datasets and is ready to be applied to deal with relatively general natural language processing tasks. On the other hand, the specialties of different engineering domains call for specific capabilities of understanding domain concepts and relationships in language processing tasks. The following questions arise: *How effective can the finetuning be in making the BERT model work for domain-specific tasks? What are the minimum finetuning requirements in terms of both training dataset sizes and the number of needed episodes? What are the effective measures that can help evaluate the performance of the domain-specific NLP tasks?* The objective here is to achieve the desired task performance with minimum finetuning efforts.

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 texts, three sentence-level datasets generated from papers in the additive manufacturing domain with different sizes are created and labeled to finetune the BERT model. The outputs

[1]Corresponding author.

Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received October 8, 2021; final manuscript received March 21, 2022; published online May 10, 2022. Assoc. Editor: Ying Liu.

of the 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 as follows:

- Investigated and assessed the effectiveness of the domain knowledge capture function of the BERT language model by collecting and creating domain-specific datasets to finetune the model and performing comparative studies for different sizes of training datasets as well as against other exiting approaches.
- 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 a summary evaluation methodology by combining machine-based and human-based evaluation methods and proposing evaluation metrics from different perspectives to uncover the information behind the summarization results.

For simplicity, this study focuses on extractive summarization and investigates the applicabilities of different BERT models. In the rest of this paper, the related work is reviewed in Sec. 2, and a systematic approach to capturing domain knowledge for a pretrained language model is described in Sec. 3. Section 4 presents a comparative study, and its results together with a comprehensive machine-and-human-based evaluation scheme. The insights obtained are discussed in Sec. 5, followed by the conclusions and future work in Sec. 6.

2 Related Work

2.1 The Development of Language Models. Manually processing massive unstructured texts and uncovering their underlying information can be time-consuming and exhausting. The evolution of natural language processing (NLP) makes it possible to reveal insights from texts. Language models, which can be considered the core of NLP techniques, are the development of probabilistic models that can learn the probability of word occurrence from documents [15].

The pioneering work in the NLP area, Word2vec, which was published in 2013, first converted text to word embedding and can be applied to many distinct downstream NLP tasks like capturing words with similar meanings [16]. However, the biggest problem of this technique is that it is not able to distinguish polysemy and cannot catch surrounding information of words [17]. To address this shortcoming, researchers have tried to use the pretraining method and directional language models to generate word representations that can dynamically change based on context [18–21]. In 2017, the transformer, a new neural network architecture, was proposed [22]. This kind of architecture can handle long-term dependencies even better than long short-term memorys (LSTMs). In order to apply the architecture to pretrain a language model and be able to be finetuned for downstream NLP tasks, a finetuning method called the Generative Pre-trained Transformer (OpenAI GPT) was introduced [20]. Later in 2019 and 2020, GPT-2 and GPT-3 were published as further developed versions of GPT [23,24]. In 2018, a Bidirectional Encoder Representations from Transformer (BERT) was proposed [12]. This language model has been pretrained using masked language model and next-sentence prediction method for assisting the model in enhancing the ability to predict the next word and handling the relationships between multiple sentences. In addition, the model was pretrained with the BooksCorpus (800M words) [25,26] and English Wikipedia (2,500M words). Moreover, it applied a bidirectional transformer architecture which can help capture both left and right contexts surrounding a word. Therefore, the BERT model can be finetuned for several downstream NLP tasks (such as text classification, question-answering, and name entity recognition) and has outperformed among NLP models [12]. In these cases, the context can be mapped to high-dimensional vector spaces and be represented by contextualized representations. Although the text generation models like GPT-2 and GPT-3 can also generate text by unsupervised finetuning process, like Zhu and Luo [25] generated design solution using GPT-2 finetuned by design problem statement, it would be difficult to collect high-quality summarization to train the generative models under this specific circumstance. Moreover, the summarizations generated by GPT-2 or GPT-3 are abstractive summarizations that can be hard to evaluate automatically and efficiently. Therefore, in this article, the BERT model is chosen as the base bidirectional language model and finetuned for extractive summarization text.

2.2 Application of Natural Language Processing Techniques in Text Summarization. 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 findings and conclusions in paperwork like clinical reports without reading the whole text [5]. Initially, the research focused on sentence features like term frequency and position of sentences in the original text, and a number of techniques have been developed [4,27]. However, these techniques may not be sufficient to capture the most significant sentences in the text and generate a high-quality summary [5,6]. Therefore, attempts have been made to extract domain knowledge from the original document, generate word presentations, and measure similar information between the words [3,6,7]. As machine learning prospered in recent years, neural network-based learning techniques have been applied to extract domain knowledge through training based on large datasets [8–11]. This 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 [3,9]. As an example, contextualized word embeddings have been widely applied to multiple downstream NLP tasks with desired performance [12,19,28]. Contextual word embeddings from the deep neural network language model have also been applied in text summarization, and the results are promising [29–32].

2.3 Application of Natural Language Processing Techniques in the Engineering Design Domain. 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 of capturing the semantic information within the text [33,34]. To address the challenge, researchers began to focus on knowledge retrieval like ontology-based retrieval to capture the ontological concepts and their relationships [35–37]. 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. As the evolution of NLP techniques enabled the capture of semantic-level knowledge from unstructured design documents, many researchers in engineering design tended to use NLP techniques to retrieve design entities and their relations to support designers. Shi et al. [38] integrated text mining approaches and unsupervised learning methods to construct a design and engineering associations ontology network. Martinez-Rodriguez et al. [39] proposed a knowledge graph construction approach to capture name entities and their binary relations from unstructured text. Sarica et al. [40,41] utilized the Word2vec method to map concepts to word embeddings and constructed a technology semantic network (TechNet) that includes scientific concepts and their semantic associations. Siddharth et al. [42] produced a large engineering knowledge graph comprising engineering

facts as <entity, relationship, entity> triples. Other than information retrieval, NLP techniques are also applied for word-level knowledge discovery in the design process. Hou et al. [43] developed an automatic way to identify and structure product affordance from a user's review using the rule-based NLP method for discovering customer needs. Han et al. [44] created a crowd-generated knowledge database to retrieve and reuse information by gathering ideas from social networking platforms to help design ideation. Han et al. [45] proposed a methodology for eliciting attribute-level user needs from online review text using the BERT model. Although word-based domain knowledge extraction can be applied to either building knowledge graphs or networks to extract the information in documents, high-dimensional sentence-level information capturing can contain more useful knowledge underlying the unstructured text. However, partly due to the lack of language models and labeled domain-specific datasets in engineering design, little work has been done on sentence-level knowledge capturing and its applications, such as text summarizations. Akay and Kim [46] introduced a sentence-level method to extract functional requirements from design documentation using the BERT model, and Ni et al. [47] presented a similarity-based approach for helping discover design solutions by utilizing a bidirectional LSTM neural network. However, only a general dataset was applied in these studies to prove the power of the language model and its potential to enhance designers' abilities. Whether sentence-level knowledge can be extracted and reused in specific tasks like text summarization remains unknown. Despite the fact that Siddharth et al. [42] provided a word-level summarizing method by selecting (entity, relationship, and entity) triples underlying unstructured text and generating knowledge graph for more condensing design-specific information, sentence-level summarization containing domain-specific information is unfinished.

2.4 Evaluation Methods for Automatic Text Summarization. Automatic text summarization has become a new trend nowadays due to the explosion of information. Researchers aim at developing various methods to condense the most significant context in the form of a summary [48]. Meanwhile, how to evaluate the performance of those summarization methods remains an urgent issue to be resolved.

The evaluation methods for text summarization can be classified into two major types [49]. One is human-based evaluation, meaning human annotators need to assess the quality of automatic summarization based on different aspects like coherence, conciseness, grammaticality, and readability [50]. Although human-based evaluation is extensive and able to provide strong support to the quality assessment, it is expensive and hard to guarantee objective results [51,52]. Therefore, researchers keep trying to propose the second type of evaluation method, machine-based evaluation methods, which can automatically evaluate the performance of different summarization methods based on a unified standard. Usually, the machine evaluation methods need to compare generated summaries with a standard summary from the original documents or experts [49]. The evaluation is often conducted based on performance metrics using recall, precision, and f-score. For example, Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which is the most popular automatic evaluation method of the last 10 years [49], uses the performance metrics to find out the overlapping content between the generated summary and the standard summary [52]. Other machine evaluation methods like bilingual evaluation understudy (BLEU) [51] and METEOR [53] are also used frequently to automatically assess the quality of summarizers.

Language models like BERT are trained with a vast amount of data and are generally useful; however, they lack specific domain understanding capabilities. Although NLP applications such as summarization have been explored in many domains, including biomedical fields, extracting and reusing sentence-level knowledge remains to be a challenge. While the NLP techniques have been applied in the engineering design area for design support and design ideation, it can be seen that the work for sentence-level text understanding is yet to be extended. Moreover, the human-based evaluation method for examining whether NLP techniques can assist experts in specific domains still needs to be explored. In this study, the domain-specific datasets were created to finetune a language model for it to acquire domain-specific knowledge and complete downstream NLP tasks. The additive manufacturing (AM) domain was chosen for this study, and the sentence-level information was captured from the relevant research papers and included in the AM datasets. Combinations of a machine evaluation method and a human evaluation method were applied to evaluate how the application of the finetuned language model can enhance the quality of text summarization in the specific domain.

3 Domain Knowledge Capture: Finetuning the Language Model

In the engineering design area, it is valuable to identify and apply the useful information or rules that underlie 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 their "understanding" in a format that can quickly and effectively help human designers grasp the essential knowledge without having to read whole lengthy documents.

In this study, a systematic approach is proposed to investigate the language models learning and capturing 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 ideas 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 contents. After applying the clustering method, corresponding summarization can finally be generated. In order to compare the performance of the language models with and without being finetuned, the same procedures after finetuning are applied to the pretrained BERT model as well. The flow of the information about this systematic approach is shown in Fig. 1.

In the following, the related details about data collection and preprocessing are illustrated in Sec. 3.1, and the structure of the language model used in this study is described in Sec. 3.2. Section 3.3 introduces the experiment method and process, followed by the evaluation methods in Sec. 3.4.

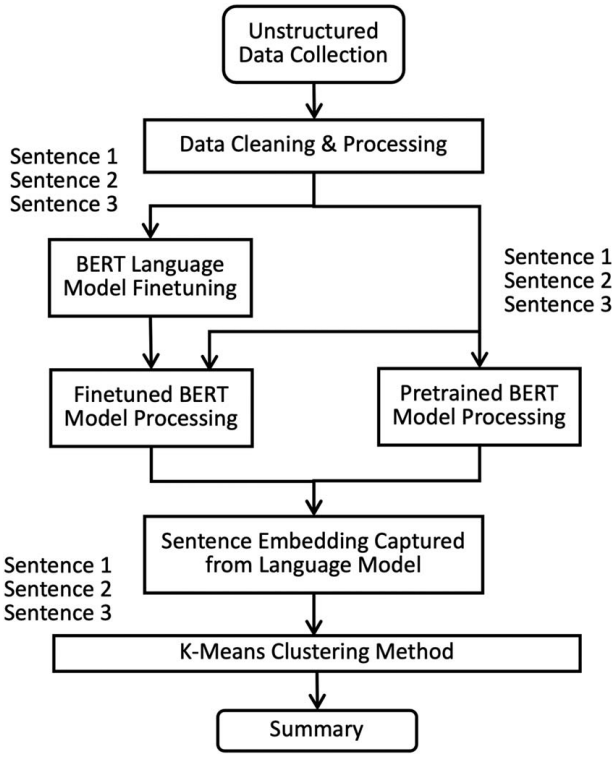

Fig. 1 The process of text summarization generation using BERT language models

3.1 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, in this research, a sample dataset is created manually to investigate 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. 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 in the additive manufacturing domain are selected from ScienceDirect and are considered original data. To train the BERT model to learn different features from the sentences that can represent the informative content in the papers, only the main parts of the selected papers are captured from original documents, including abstract, introduction, and conclusion sections, which is under the assumption that these sections including the most significant content also contain rich domain-specific knowledge. Due to the requirements of finetuning the BERT model, the paragraphs need to be tokenized up into individual sentences using the NLTK toolkit [54] (e.g., "Finishing of components originating from additive manufacturing (AM) is critically important for providing them with adequate tolerances and fatigue life. Based on 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." to "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."). Finally, there are 505, 2020, and 6167 sentences correspondingly tokenized from 38, 60, and 172 most recent papers 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 those circumstances, the sentences inside the tables are extracted by manual work. In addition, some irrelevant sentences like figure captions are removed.

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 labeled 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 the abstract and conclusion parts which display the main ideas and findings of the paper are all labeled as 1, while sentences in the introduction part which convey information about related background would be considered as less important sentences and are all labeled as 0. Under an ideal circumstance, multiple operators ought to label all the sentences in the datasets so that the accuracy and reliability of labeling results can be measured. However, due to resource limitations, the labeling results were cross-checked by two operators. For each paper, the labeling results generated from the first operator would be examined by the second operator; and no significant differences were found. 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.

3.2 The BERT Language Model. BERT [12] is a language representation model developed by Google. This new model is different from past language models like recurrent neural networks (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. First, 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. The transformer contains two mechanisms: an encoder reads the text input, and a decoder output the prediction of the tasks [22]. Since the attention mechanism contains multiple attention heads, it can significantly enhance the computational performance and increase training accuracy even with small datasets.

In addition, the input representation of the BERT model can be constructed in 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. In this study, the BERT model is used for converting words and sentences into contextualized vectors and completing the summarization tasks. Different downstream language tasks (Question answering, Sequence classification, Name entity recognition, etc.) can be achieved with great performance based on the same architecture as Fig. 2 [12] below shows.

In this research, in order to complete the finetuning process for the sentence classification task, different factors and

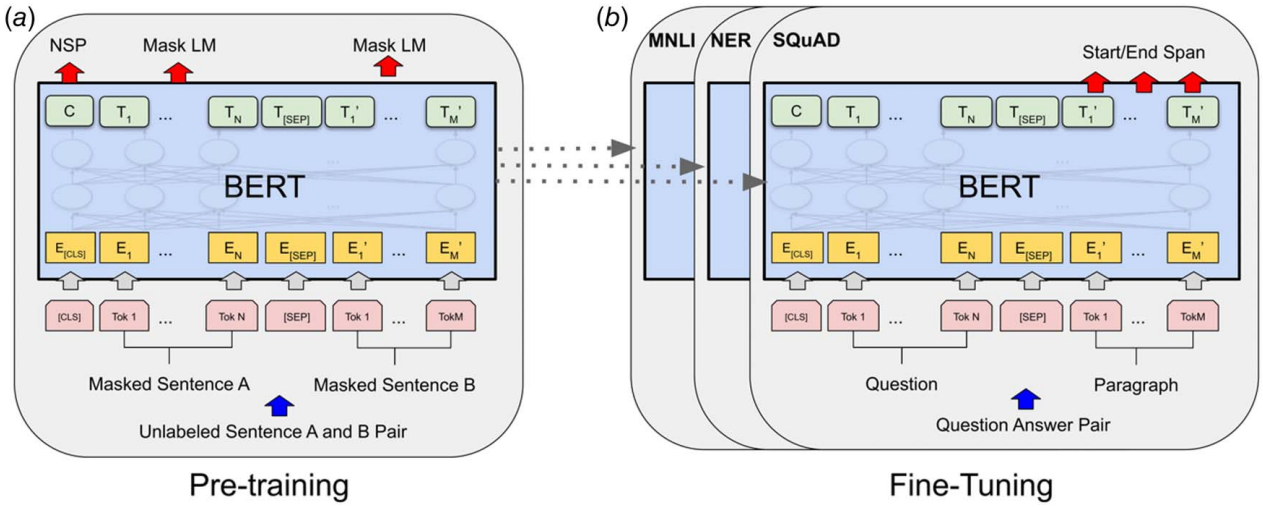

Fig. 2 The architecture of (*a*) pretrained BERT model and (*b*) finetuned BERT model

hyperparameters need to be set to help the BERT model achieve the best performance:

(1) *Preprocessing of long sentences*. The maximum sequence length of the BERT model was set to 256 and then padded with zeros.
(2) *Selection of layers*. There are many different pretrained BERT models provided by the Google AI (artificial intellligence) platform. In this study, an uncased BERT-base model consisting of an embedding layer, 12 encoding layers, and a pooling layer with 768 hidden sizes is selected as the pretrained model. Two additional fully connected layers for flattening the output and adding dropouts and a SoftMax activation function layer were added for performing the text classification task to predict the labels out of {0, 1}. Figure 3 below displays the network architecture for the sentence selection/sequence classification task.
(3) *Selection of optimizer*. Adam optimizer was used with a 2×10^{-5} learning rate in the finetuning process since this optimizer is an optimization algorithm for stochastic gradient descent during deep learning model training and can be applied to deal with sparse gradients on the noisy problem [55].

4 Hyperparameter setting. The authors of the BERT model recommend using 2–4 epochs to train the BERT model [12]. In this study, four epochs are chosen to finetune the BERT model. Moreover, the learning rate is 2×10^{-5}, and the batch size is set to 16.

3.3 Experimental Method. In this study, the BERT model and a *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 (as circles shown in Fig. 4 below), and the sentences that are nearest to the centroids of clusters will be selected and be included in the final summarization. The specific process of captured sentence embeddings and *k*-means clustering are shown in Fig. 4 below.

After obtaining the finetuned BERT model, sentence embeddings can be captured from the network architecture. Although Sentence-BERT, or SBERT, has been applied to directly generate sentence embeddings [56], from a finetuning point of view, the requirements for designing domain-specific training datasets are rather complicated, making it difficult to finetune SBERT. Therefore, in this study, a sentence embedding generation method with a simple dataset design is applied. In the method, 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 $N \times E$ vectors where N represents the number of sentences and E is 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, the *k*-means method [57] 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 is 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 is investigated.

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

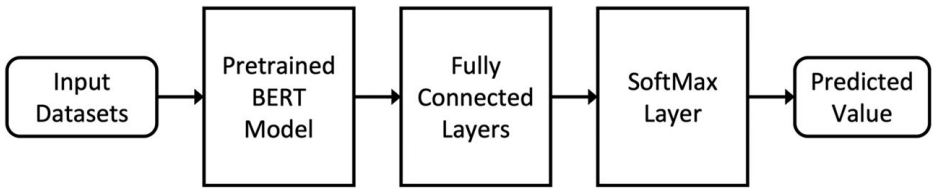

Fig. 3 The network architecture of the BERT model for sentence classification tasks

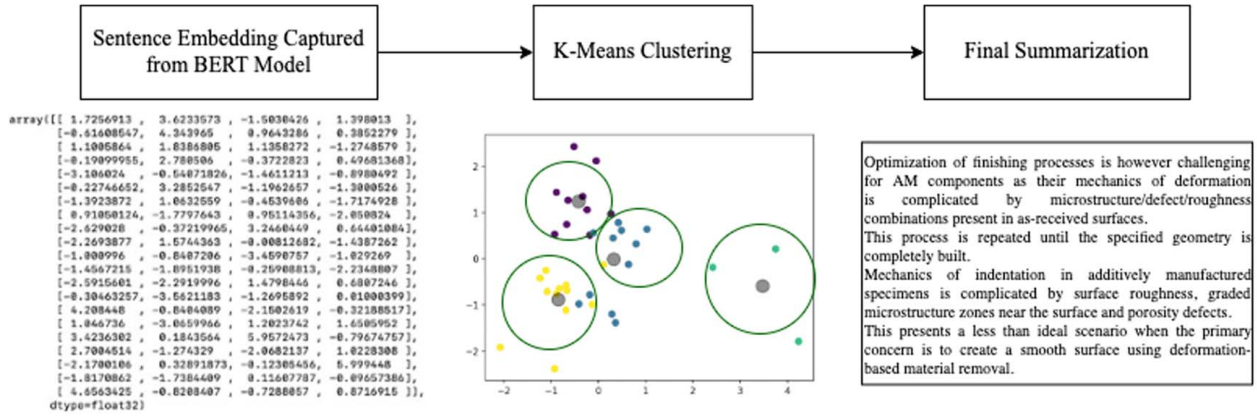

Fig. 4 Examples of sentence embedding and *k*-means clustering

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 [58] 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 [59] 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 [60] is a topic-based approach that evaluates the significance of sentences by their singular value decomposition (SVD) values. Random baseline, which selects sentences in the original text randomly, is also applied and 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, both machine evaluation methods and a human evaluation method are chosen for evaluating the performance of summarization systems in this paper. ROUGE [52] is a widely used intrinsic evaluation due to its efficacy. In ROUGE, precision, recall, and F-score are applied as evaluation metrics to evaluate the quality of the summary. It generates this evaluation metric by comparing the standard summary, or reference summary, and the automatically generated summary. Based on different criteria, it measures the overlapping information between the reference summary and the 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 the 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 single-document summary cases [61].

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 [62]. Commonly, researchers may use a human-made summary or abstract of papers as a 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. 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 [63], keywords in a document can represent the most significant idea of its content, meaning that a summarization would contain more high-frequent words. Consequently, in addition to ROUGE, the word-frequency measurement, which is a statistical-based method, is also considered an evaluation method to measure the quality of the generated summaries based on the assumption that the more frequent terms in a document are more important and more indicative of the topic [63]. 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 by long-length sentences, the score of each sentence will be divided by the number of words in the sentence.

The automatic evaluation using the ROUGE method and word-frequency algorithm presented previously has its limitation since it is difficult to discover whether the summary can help AM researchers in the real world and whether useful domain information has been captured. Researchers doubted that ROUGE could be misleading when it is the only method used for evaluation [64]. Also, a study about the correlation between ROUGE and human evaluation shows, in general, the correlation is low [65]. Therefore, eliciting human judgment becomes indispensable to investigate the influence of summarizers on the summaries and the usefulness of these summaries to AM researchers. Hence, a human evaluation-based user study is performed to measure the performance of three different summarizers, namely, the pretrained BERT model, the finetuned BERT model with the largest dataset, and the Text Rank.

4 Results and Comparison Study

During the experiment described earlier, the BERT model is finetuned using different sizes of domain-specific datasets. The test and validation of the performance of the finetuned BERT models are evaluated using both machine evaluation methods and human evaluation methods. Currently, researchers have studied the capability of different language models and their finetuning performance. For example, Ethayarajh [66] compared the word embedding generated by BERT, ELMo, and GPT-2 and found that the contextualized embedding can contain more information compared to static embedding; Nguyen et al. [67] compared the finetuning results from different combinations of classification layer with BERT model. The main purpose of this paper is to investigate the properties of finetuning on a well-trained language model with the objective of achieving the desired performance with minimum training efforts. Therefore, only text summarizations generated by the BERT-based model are compared. In the following subsections, the details of the summarization results and the comparison study are presented.

4.1 Finetuning Results. When the size of the training dataset is different, the finetuned BERT model can show different results,

as one may expect. An interesting question is how the positive effect of increasing size may diminish. In order to assess the influence of the size of training datasets on the BERT training results, different datasets are created with [500, 1000, 1500, 2000, 3000, 4000, 5000, 6000] sentences, respectively. These datasets are used to train 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 applied to measure the testing accuracy of models.

In order to investigate the relationship between the size of training datasets and training accuracy as well as testing accuracy, one experiment of finetuning the BERT model with different-sized training datasets was conducted. Since the learning algorithm behind the BERT model is stochastic, a difference exists in the performance of finetuning process across several runs. For summarizing the performance, the BERT model was finetuned and evaluated with the same training and testing dataset under the same hyperparameter settings 10 times, and the accuracy of each run was recorded. The mean of the accuracy of the finetuned model will be calculated and used to represent the mean performance of the finetuning models, and the standard deviation of it will be used to represent the variance.

Table 1 shows the mean value and standard deviation of the training accuracy and testing accuracy for different BERT models. For the BERT model finetuned by 505 sentences, as Table 1 shows, the testing accuracy of it is about 57.5%. As the size of the dataset increases, the testing accuracy of BERT models is also improving from 57.5% to 77.5%. 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.

Figure 5 depicts the trend of training accuracy and testing accuracy changing with the size of the training dataset. It can be seen that there is a sharp performance increase as the size changes from 500 sentences to 1000 sentences. After that, the performance increase slows down. In addition, as the size of datasets expands, the difference between training accuracy and testing accuracy becomes smaller.

Figure 5 shows that 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 1000 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 even with a 1000-sentences increment can only result in moderate accuracy gains. From a perspective of minimizing training effort for finetuning, Fig. 5 provides a clear indication of the importance of the size of the 1000-sentence dataset. Further increasing the dataset size may not contribute to the task performance of the BERT model. There exists a "minimum finetuning size (MFS)" for effective BERT performance.

Moreover, as the size of the dataset increased to 6000 sentences and beyond, the accuracy improvement almost stalled. Nevertheless, as Table 1 shows, the finetuned BERT model with 6000 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, although the knowledge learned after the "minimum finetuning size" is relatively limited depending on the application tasks.

Table 1 Training and testing accuracy of different BERT models

	Finetuned with various # of sentences							
	500	1000	1500	2000	3000	4000	5000	6000
Training	0.694 ± 0.003	0.786 ± 0.013	0.809 ± 0.014	0.821 ± 0.017	0.828 ± 0.019	0.835 ± 0.021	0.856 ± 0.022	0.890 ± 0.021
Testing	0.575 ± 0.013	0.743 ± 0.026	0.762 ± 0.017	0.776 ± 0.032	0.806 ± 0.023	0.823 ± 0.020	0.842 ± 0.016	0.854 ± 0.013

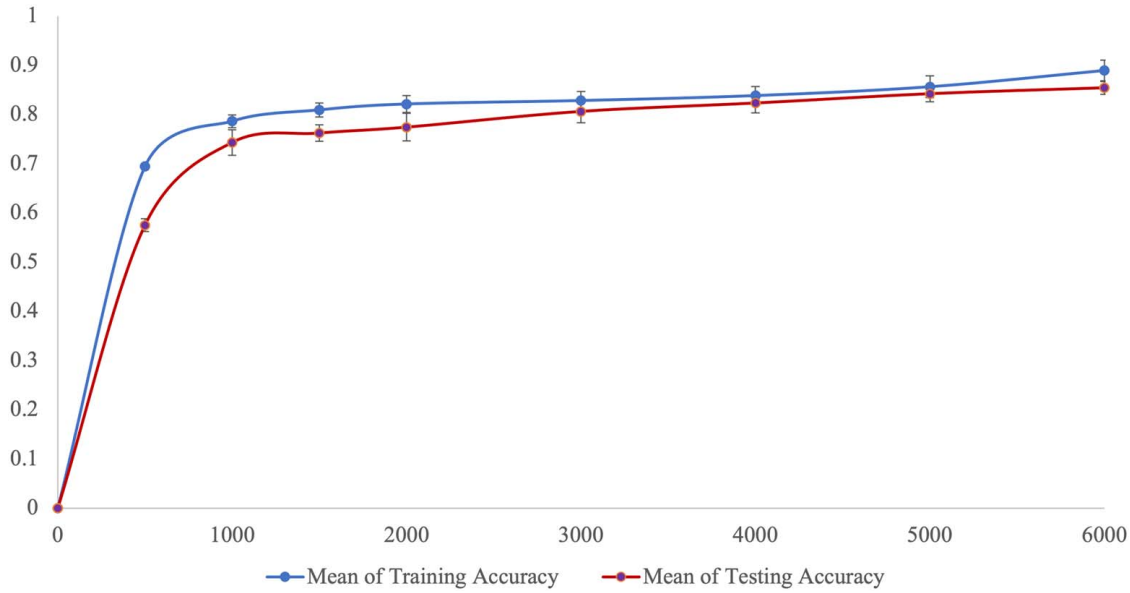

Fig. 5 Training and testing accuracy changing with different sizes of datasets

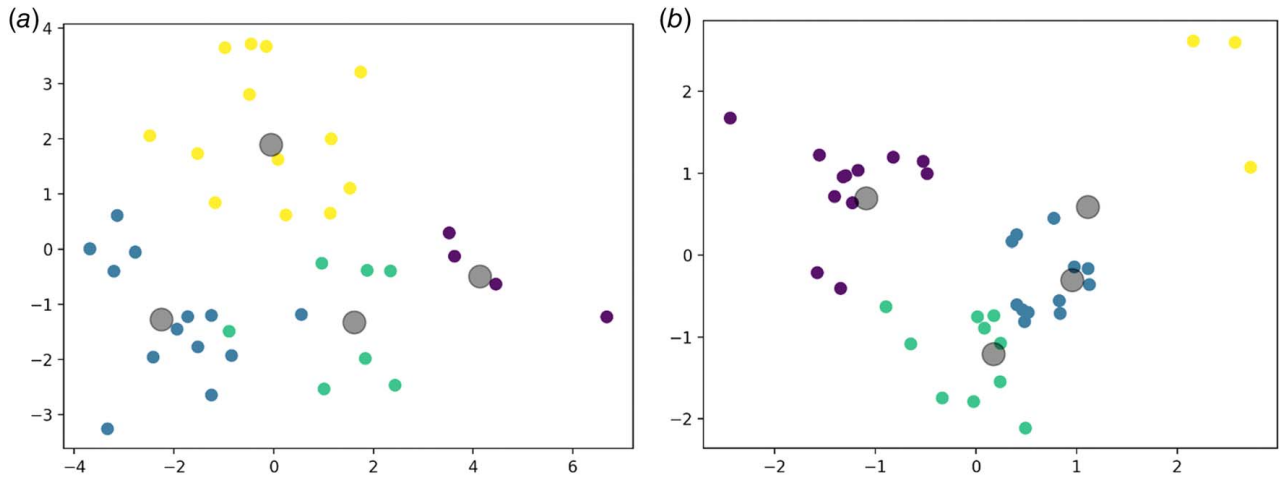

Fig. 6 Representation of 2D sentence embeddings captured from (*a*) pretrained model and (*b*) finetuned model by 500 sentences

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 two-dimensional (2D) coordinate. Based on the plot, it can be seen that the sentence embedding has significantly changed after the model is finetuned.

An illustrative example is provided below to demonstrate the results generated by BERT-based summarizers from the same paper. From the example, it shows that the summary by the finetuned BERT model has denser information and better structure than the summary generated by the pretrained BERT model.

Example 1: Pretrained BERT-based summarizer

"However, with advances in materials, process controls, and robotics, the field of additive manufacturing continues to progress towards becoming a viable option for large-scale, high-volume industries. Metal AM processes can be classified as powder bed or wire/ blown powder feed processes. An electric arc forms between the wire and the substrate, which melts the wire and deposits a bead of molten metal along the predetermined path. Greer et al. discussed the design rules for MBAAM and demonstrated the fabrication of a large-scale excavator arm to highlight the potential use of the system in the service and repair industry, owing to its high deposition rates in conjunction with its near-net shaping capability. They reported scattering in elongation that they attributed to microstructural heterogeneity and discontinuities such as local soft spots. The use of AM techniques can drastically increase productivity and reduce the lead time and cost of manufacturing molds and dies. In a survey of 96% of mold manufacturers, it was found that they have plans to utilize AM for overcoming these obstacles, especially since customer demands are continuing to put pressure on mold builders. ORNL, Wolf Robotics and Lincoln Electric have developed an MBAAM system that uses high deposition rates and low-cost wire feedstock material that can be used to manufacture molds and dies for the composite industry."

Example 2: Finetuned with the largest dataset BERT-based summarizer

"Historically, AM has not been suitable for high-volume production or large-scale projects. These systems are used with a wide array of materials/alloys. The MBAAM system was developed by Oak Ridge National Laboratory (ORNL) in a collaboration with Wolf Robotics and Lincoln Electric. Greer et al. discussed the design rules for MBAAM and demonstrated the fabrication of a large-scale excavator arm to highlight the potential use of the system in the service and repair industry, owing to its high deposition rates in conjunction with its near-net shaping capability. This paper also discusses the properties and the related microstructure of AM parts manufactured by the MBAAM system using mild steel wire ER70S-6. ORNL, Wolf Robotics and Lincoln Electric have developed an MBAAM system that uses high deposition rates and low-cost wire feedstock material that can be used to manufacture molds and dies for the composite industry. The system can be effectively used to create a part with an overhang angle of 90 degrees. The mechanical properties of the printed AM structure were found to be planar isotropic in nature, which is crucial for mold and die applications."

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 the pretrained model. More detailed experiments and evaluations are discussed below.

4.2 Testing Data Preparation in Summarization Evaluation. According to what Lin [61] 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 conclusion sections are extracted for generating summarization, while the abstracts of papers are used as the 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 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 4, which can represent the differences in the performance of these summarization methods.

4.3 Machine Evaluation

4.3.1 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 second, it is applied to compare ROUGE scores of summaries generated by the pretrained BERT model and finetuned BERT models.

Table 2 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 is 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 is 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 is 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.

4.3.1.1 Parameterization. In order to identify proper parameter settings for achieving the best performance, the impact of different sizes of dimensions for sentence embeddings is investigated. As shown in Table 2, the summaries generated by sentence embeddings with different dimension sizes are distinct, indicating that dimensionality reduction, which 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 needs to be investigated. In this experiment, to evaluate the quality of summaries under different dimension sizes, ROUGE evaluation is applied to measure the impact 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 settings to be no larger than the number of samples. Since the minimum number of samples in the dataset is 22, the components are set to less than 22 dimensions. Table 3 below shows the mean scores of summaries with different dimensions as well as the underlying language models.

According to the results shown in Table 3, the loss of performance exists with dimensionality reduction. Specifically, sentence embeddings with 768D can generate summaries with higher scores, meaning that the 768-dimension sentence embedding model captures more information and has a better performance compared to other dimensional models. In other words, high-dimensional models can acquire the most representative information from the

Table 3 Mean value under different dimensions

	Pretrained BERT			Finetuned BERT (500)			Finetuned BERT (2021)			Finetuned BERT (6067)		
Dimensions	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**

Note: The number in brackets represents the size of dataset.
Boldfaced values represent the best performance among all the testing cases.

Table 4 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

Note: The number in brackets represents the size of the dataset.
Boldfaced values represent the best performance among all the testing cases.

papers. Therefore, in this experiment, all the summaries are generated by 768D sentence embeddings in order to ensure the best results.

4.3.1.2 Comparisons among different summarizers. Table 4 presents the mean value of ROUGE-1, ROUGE-2, and ROUGE-L scores acquired from the BERT-based summarizers and other non-BERT-based approaches. As Table 4 shows, the performance of the finetuned BERT model trained by 6167 sentences exceeds other BERT-based summarizers and non-BERT-based summarizers with regard to ROUGE-1, ROUGE-2, and ROUGE-L scores significantly based on the ANOVA test ($p < 0.01$).

Taking the 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, compared 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 the other text summarizers, while the recall scores of finetuned BERT models are relatively low or indifferent. In addition, considering the comparisons among the three finetuned BERT models with different dataset sizes, the mean scores increased as the size of datasets increased, which is also consistent with the corresponding accuracy of the BERT training models discussed above. For instance, when the size of the dataset was enlarged 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 compared to the pretrained BERT model.

4.3.2 Statistical Evaluation. Table 5 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 the finetuned BERT models is enhanced. Specifically, the mean scores of the summarizers with a larger dataset can be 14.50% higher than others. A closer observation indicates that most of the 14.50% increase was achieved through finetuning based on the 500-sentence datasets, reflecting the "minimum finetuning size" discussed in Sec. 4.1.

Table 5 Average word-frequency score of summarizations generated by different BERT models

Summarizers	Mean	Min	Max
Pretrained BERT	3.432 ± 0.48	2.475	4.660
Finetuned BERT (500)	3.785 ± 0.82	2.175	5.572
Finetuned BERT (2012)	3.790 ± 0.85	2.414	5.817
Finetuned BERT (6167)	3.930 ± 0.76	2.524	5.579

*Note: The number in brackets represents the size of the dataset.

4.4 Human Evaluation. A human evaluation user study is performed to evaluate whether AM researchers can have different impressions on summaries generated by different summarizers and whether the summaries generated by the finetuned BERT models can contain more important domain-specific information and assist AM researchers with a certain background. In this user study, three summarizers are selected based on the machine-based automation evaluation results. The pretrained BERT model and the finetuned BERT model with the largest dataset are chosen due to their high-performance quality in summary generation. Meanwhile, Text Rank, which is a summarizer with the best performance among all the non-BERT-based summarizers, was also selected to be compared. A pairwise comparison among these three summarizers is then conducted. This human evaluation user study is carried out in the form of online surveys hosted using Google Forms. In the end, 11 students and professors with additive manufacturing backgrounds recruited via university mailing lists as well as online social networks have participated in this online survey. Each participant evaluated groups of summaries from 10 randomly selected papers from the additive manufacturing area. The participants are anonymous, and only their responses are collected. The average time of the survey completion is about 40 min.

4.4.1 Design of Online Survey. The online survey contains 10 groups of summaries generated from the three summarizers mentioned previously on the same sets of additive manufacturing papers. It was designed for measuring the quality of summaries generated from different systems, i.e., Pretrained BERT model, Finetuned BERT model, and Text Rank. In order to make it convenient for participants, the online survey was partitioned into 2 parts, and each part contained five groups of data. In each group, the procedures of human evaluation are as follows.

Step 1: To help participants quickly grasp the main content and figure out the focus of a paper, the abstract of the paper will be shown in the survey first. Figure 7 presents an example of the first step in the survey. Participants need to read through the abstract at the beginning and then start to compare the groups of summaries based on their understanding.

Step 2: After having some understanding of the main idea of the paper, participants are asked to perform comparisons among the summaries generated from three different summarizers without knowing which summarizer generated which summary. The summaries for one paper are grouped and ordered randomly, as shown in Fig. 8. Participants need to

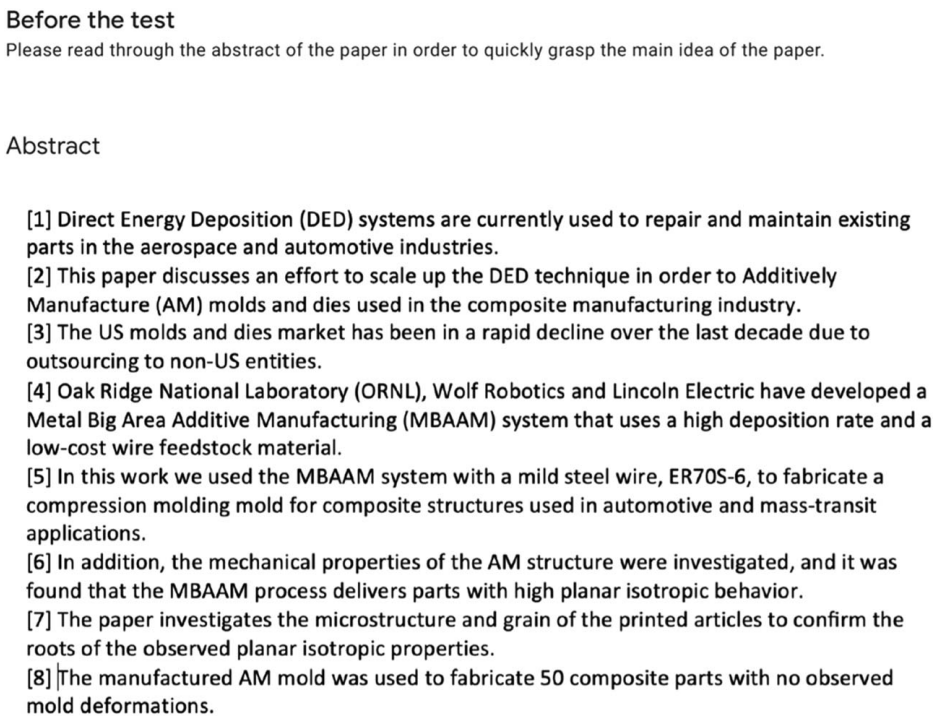

Fig. 7 The abstract of one paper displayed in the first step

answer questions from distinct aspects based on the content of the summaries.

Step 3: Participants are asked to answer questions based on three different aspects of the summaries that measure the performance of the summarizers. The questions are selected and used according to [65]. For each group of summaries, participants are asked to rank them by selecting options from the best to worst according to their understanding, as shown in Fig. 9. For each rank, i.e., best, good, or worst, only one summary can be selected. The details of the questions are presented below:

S1: Which one of them is more informative about additive manufacturing?

S2: Which one of them expresses the meaning closest to the abstract?

S3: Which one of them has the least redundant information?

These questions are used to evaluate the summaries from different perspectives. Specifically, S1 relates to Informative Coverage, S2 relates to Informative Relevance, and S3 is about Informative Redundancy. These three aspects can be used to clarify whether the summaries generated by different summarizers can contain important information and be applied to help AM researchers.

4.4.2 Results of Human Evaluations and Findings. Tables 6–8 show how often the participants ranked each summarizer from different aspects. In the table, data in each column mean the percentage of participants voting for the summarizers. A one-way ANOVA test was carried out for every pair of summarizers in order to assess the significant difference among the summarizers and to check whether the results were caused by chance. Finally, the ANOVA test shows there is a significant difference ($p < 0.05$) between the pretrained BERT model, the finetuned BERT model as well as the Text Rank in terms of information coverage, information relevancy, and information redundancy, indicating that the results were not caused by chance.

Table 6 shows the results regarding question 1, which relates to informative coverage. According to Table 6, most of the evaluation results about summaries generated by the BERT-based model lay in the range of good quality to best quality. About 81.82% of the evaluators think the summaries generated by the pretrained BERT model are above good quality in the aspect of informative coverage. About 61.18% of evaluators think the summaries generated by the finetuned BERT model are above good quality.

However, the summary generated by Text Rank gained polarized results. 41.81% of the evaluators think Text Rank can generate the most informative summary, while 43.64% of them think it is the worst. It can be speculated that Text Rank tends to select longer-length sentences in the summary due to its specific algorithm. Some participants think longer paragraphs can contain more information, while others think most of the information in the summary is redundant.

Table 7 shows the results relating to informative relevance. From the table, it can be seen that 42.73% of the evaluators think the finetuned BERT model can generate summaries that express the closest meaning to the abstract. And 45.45% of the evaluators think Text Rank generates the worst results. Also, the pretrained BERT model has better performance than Text Rank since the Text Rank tends to capture longer sentences without considering the meaning. The results imply that after being finetuned, the BERT model did learn how to capture the significant sentences and include the sentences in the final summary. Therefore, it earns the best results compared to the other two summarizers.

Table 8 shows the results in terms of informative redundancy. From the table, the results show that the finetuned BERT model achieved the best performance on informative redundancy; 46.4% of the participants think finetuned BERT model can generate the least redundant summaries. Meanwhile, the pretrained BERT model is ranked second, and Text Rank is ranked as the worst summarizer under this category.

The results about informative redundancy can also support the results of informative coverage as well as informative relevance. The evaluators think summaries generated by the finetuned BERT model contain the least redundant information while the summary generated by Text Rank is the most redundant.

From the earlier results, the finetuned BERT model can be considered as the best summarizer among the three summarizers. Comparing the finetuned BERT model with the pretrained BERT model,

(*a*)	Summary A	[1] However, with advances in materials, process controls, and robotics, the field of additive manufacturing continues to progress towards becoming a viable option for large-scale, high volume industries. [2] Metal AM processes can be classified as powder bed or wire/ blown powder feed processes. [3] An electric arc forms between the wire and the substrate, which melts the wire and deposits a bead of molten metal along the predetermined path. [4] Greer et al. discussed the design rules for MBAAM and demonstrated the fabrication of a large scale excavator arm to highlight the potential use of the system in the service and repair industry, owing to its high deposition rates in conjunction with its near net shaping capability. [5] They reported scatter in elongation that they attributed to microstructural heterogeneity and discontinuities such as local soft spots. [6] The use of AM techniques can drastically increase the productivity and reduce the lead time and cost of manufacturing molds and dies. [7] In a survey of 96% of mold manufacturers, it was found that they have plans to utilize AM for overcoming these obstacles, especially since customer demands are continuing to put pressure on mold builders. [8] ORNL, Wolf Robotics and Lincoln Electric have developed an MBAAM system that uses high deposition rates and low-cost wire feedstock material that can be used to manufacture molds and dies for the composite industry.
(*b*)	Summary B	[1] Historically, AM has not been suitable for high-volume production or large-scale projects. [2] These systems are used with a wide array of materials/alloys. [3] The MBAAM system was developed by Oak Ridge National Laboratory (ORNL) in a collaboration with Wolf Robotics and Lincoln Electric. [4] Greer et al. discussed the design rules for MBAAM and demonstrated the fabrication of a large scale excavator arm to highlight the potential use of the system in the service and repair industry, owing to its high deposition rates in conjunction with its near net shaping capability. [5] This paper also discusses the properties and the related microstructure of AM parts manufactured by the MBAAM system using mild steel wire ER70S-6. [6] ORNL, Wolf Robotics and Lincoln Electric have developed an MBAAM system that uses high deposition rates and low-cost wire feedstock material that can be used to manufacture molds and dies for the composite industry. [7] The system can be effectively used to create a part with an overhang angle of 90 degrees. [8] The mechanical properties of the printed AM structure were found to be planar isotropic in nature, which is crucial for mold and die applications.
(*c*)	Summary C	[1] At ORNL, we developed the MBAAM system to fill the need for an AM system for large, fast, and competitive cost production of molds and dies for the composite industry. [2] Greer et al. discussed the design rules for MBAAM and demonstrated the fabrication of a large scale excavator arm to highlight the potential use of the system in the service and repair industry, owing to its high deposition rates in conjunction with its near net shaping capability. [3] ORNL, Wolf Robotics and Lincoln Electric have developed an MBAAM system that uses high deposition rates and low-cost wire feedstock material that can be used to manufacture molds and dies for the composite industry. [4] MBAAM is fundamentally similar to gas metal arc welding (GMAW) but is equipped with a closed loop control system to ensure geometric accuracy over the course of the build using a correction-based approach to overcome the dynamic nature of welding. [5] 11% of mold manufacturers aid their plan to use advantages of AM on conventional molds to create such complex features with less manufacturing steps and cost. [6] However, there is limited work on demonstrating the suitability of the system for fabricating low cost molds and dies. [7] This reduction in weight will result in an enhancement of the overall fuel efficiency of a vehicle and allow the original equipment manufacturer (OEM) to comply with the corporate average fuel economy (CAFE) standards and regulations. [8] In a survey of 96% of mold manufacturers, it was found that they have plans to utilize AM for overcoming these obstacles, especially since customer demands are continuing to put pressure on mold builders.

Fig. 8 An example of comparison among summaries generated by three summarizers: (*a*) pretrained BERT model, (*b*) finetuned BERT model, and (*c*) text rank

1. Which one of them is more informative about additive manufacturing? *

Mark only one oval per row.

	Summary A	Summary B	Summary C
Best	◯	◯	◯
Good	◯	◯	◯
Worst	◯	◯	◯

Fig. 9 Example of ranking options of content evaluation

Table 6 The data analysis results of informative coverage

Summarizers	Best	Good	Worst
Pretrained BERT model	0.364 (40/110)	0.455 (50/110)	0.182 (20/110)
Finetuned BERT model	0.218 (24/110)	0.4 (44/110)	0.382 (42/110)
Text Rank	0.418 (46/110)	0.145 (16/110)	0.436 (48/110)

Note: Data in the bracket presents the number of votes. The differences between each pair are significant ($p < 0.05$) based on the one-way ANOVA result.

Table 7 The data analysis results of informative relevance

Summarizers	Best	Good	Worst
Pretrained BERT model	0.264 (29/110)	0.427 (47/110)	0.309 (34/110)
Finetuned BERT model	0.427 (47/110)	0.337 (37/100)	0.236 (26/110)
Text rank	0.309 (34/110)	0.236 (26/100)	0.455 (50/110)

Note: Data in the bracket presents the number of votes. The differences between each pair are significant ($p < 0.05$) based on the one-way ANOVA result.

Table 8 The data analysis results of informative redundancy

Summarizers	Best	Good	Worst
Pretrained BERT model	0.3 (33/110)	0.464 (51/110)	0.236 (26/110)
Finetuned BERT model	0.464 (51/110)	0.309 (34/110)	0.227 (25/110)
Text rank	0.236 (26/110)	0.227 (25/110)	0.536 (59/110)

Note: Data in the bracket presents the number of votes. The differences between each pair are significant ($p < 0.05$) based on the one-way ANOVA result.

although the results of the pretrained BERT summarizer about information coverage are better than that of the finetuned BERT model, according to the results from information relevance and redundant, it can be seen that much information contained in the summaries generated by the pretrained BERT model is related to additive manufacturing but not the focus of the paper. Meanwhile, the summaries generated by the finetuned BERT model include more overlapping information with the abstract and less redundant information about additive manufacturing, which can be considered as representing the main idea of the papers. Moreover, the polarized results of Text Rank in information coverage can be explained. The summaries generated by Text Rank have a longer length and more information about additive manufacturing compared to the other two summarizers, but they do not contain the most significant information of the original papers.

The results of the human evaluation indicate that the summaries generated by the finetuned BERT model can capture the main idea and express the most significant meaning with the least redundant sentences. After being finetuned with a domain-specific dataset containing additive manufacturing information, the finetuned BERT model is able to create the summaries including the closest content to the abstract compared to the other two summarizers.

Overall, the finetuned BERT summarizer has the best performance compared to the pretrained BERT model and Text Rank. After carrying out the ANOVA test, it is shown that there is a significant difference ($p < 0.05$) between each pair of summarizers. Therefore, according to human-based evaluation, AM researchers present their agreement to the summaries generated by the finetuned BERT model based on their domain knowledge. This implies that the finetuning process can enhance the possibility of a language model to capture the domain information and utilize that information in various applications like text summarization.

5 Discussion

5.1 Summarization 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 result demonstrates that the sentence embeddings with higher dimensions can capture more information about the text documents.

5.2 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 are 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, which indicates that BERT-based summarizers are able to capture more significant sentences in the original texts.

5.3 Domain Knowledge Capture Through Model Finetuning. Comparing the pretrained BERT model and the finetuned BERT models, the results indicate that the BERT model can increase the number of keywords in summaries after being finetuned by the domain-specific datasets. Moreover, based on the results from the 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 better than other summarizers due to the additional and domain-specific training. Besides, in comparing recall scores and precision scores of BERT-based summarizers, the result shows that the recall scores decreased while the precision scores were highly enhanced after the finetuning process. The high precision score means the summaries generated by finetuned BERT models have more overlapping words with the original abstract compared to other models. Meanwhile, the low recall score illustrates that much information in the original text is redundant from the perspective of the finetuned BERT models. The results of human evaluation also provided strong evidence that the BERT model is capable of capturing useful domain knowledge through finetuning process. That domain information can be utilized for further design support.

5.4 Performance of Sentence Representations From Various Sizes of Data Set. Comparing the finetuned BERT models trained with domain datasets of distinct sizes, the model trained by larger datasets outperforms those by smaller ones, which is not surprising. According to the ROUGE scores and the statistical scores, the finetuned BERT models with the largest dataset always achieve the best performance. However, as indicated by Fig. 5 and through closer observations of the results shown in Tables 4 and 5, the increase in the performance with respect to the increase of training dataset size is rather nonlinear. The significant increase happens when the size varies from 0 (e.g., no finetuning) to 500 and then to 1000. After that point, the performance gain is gradually diminishing. This result is significant in that the language models like BERT that have been pretrained with a vast range of datasets have a great potential to be customized for

domain-specific tasks with only a minimum finetuning effort. No significant datasets or lengthy training time is needed.

6 Conclusions and Future Work

In this study, sentence embeddings that convert the unstructured texts to multidimensional vectors are extracted by applying BERT models and then used through a *k*-means method to capture the main idea of different papers and generate paper summaries. The evaluation results of the BERT models, together with other non-BERT models, demonstrate that the word representations in finetuned BERT models can capture the informative context of the papers effectively.

Based on the results and discussions described earlier, it can be concluded that contextual embeddings can enhance the performance in NLP tasks like text summarization. In addition, the 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 contain semantic and contextual information and have a great potential for processing other NLP tasks in different domains. From an engineering support point of view, the high effectiveness of the finetuned BERT models has opened ways to developing extensive NLP tools to support engineering knowledge capture, personal NLP-based design assistance, and engineering collaboration. An important advantage of this approach is that the language models like BERT that are pretrained with vast datasets can be customized for domain-specific tasks without requiring vast finetuning datasets and lengthy training times thanks to the existence of the relatively small minimum finetuning dataset size.

The proposed bidirectional language model-based approach has several limitations. First, sentence embeddings based on averages of word embeddings are easy to implement but may lead to loss of information. Alternative methods, including Universal Sentence Encoder [68], Sentence-BERT [56], and InferSent [69], need to be investigated for possible performance enhancement. Second, the *k*-means clustering for summarization is an easy to interpret method but may not handle high-dimensional vectors as well as other methods such as t-distributed stochastic neighbor embedding (t-SNE) [70]. Third, the results and insights obtained thus far are based on the BERT model. Further work is needed to explore other language models like GPT-2 and GPT-3. Lastly, the manual labeling process needs to be further elaborated. One potential machine-based approach is to take all the sentences in the abstract as label 1 and those in the introduction as 0. Another direction is to devise some unsupervised learning methods as those used with GPT-2.

In this paper, the contextualized embeddings are only used for general summarization through domain-specific finetuning. Future work includes exploring the features of the sentence embeddings, examining the clustering properties, and going beyond summarization. Besides optimizing the experimental methods and applying alternative language models to enhance the performance, design-perspective information capturing (e.g., customer needs, functional requirement, design solutions, constraints, and problem-solving process), including intra- and inter-perspective structures, and design-specific summarizations, will be conducted to assist designers. The long-term goal is to realize highly "intimate" computer-aided design by using BERT-like language models to augment design engineers' working and thinking processes based on vastly available documents.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request. The data and information that support the findings of this article are freely available at: Contact the corresponding author.

Appendix A

Table 9 The standard deviations under different dimensions

	Pretrained BERT			Finetuned BERT (500)			Finetuned BERT (2021)			Finetuned BERT (6067)		
Dimensions	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	0.04	0.02	0.03	0.04	0.03	0.04	0.03	0.02	0.03	0.03	0.02	0.03
4D	0.03	0.02	0.03	0.04	0.03	0.04	0.04	0.02	0.04	0.04	0.02	0.04
20D	0.03	0.02	0.05	0.04	0.03	0.05	0.03	0.02	0.03	0.03	0.02	0.03
768D	0.03	0.02	0.03	0.04	0.02	0.05	0.04	0.02	0.05	0.03	0.02	0.03

Note: The number in brackets represents the size of dataset.

Appendix B

Table 10 Standard deviations 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	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
Finetuned BERT (500)	0.04	0.04	0.04	0.02	0.02	0.02	0.05	0.05	0.05
Finetuned BERT (2012)	0.04	0.04	0.04	0.02	0.02	0.02	0.05	0.05	0.05
Finetuned BERT (6167)	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
Text rank	0.05	0.05	0.04	0.02	0.02	0.02	0.05	0.05	0.05
KL-Sum	0.05	0.05	0.05	0.03	0.03	0.03	0.05	0.05	0.05
LSA	0.05	0.05	0.05	0.03	0.03	0.03	0.05	0.05	0.05
Random Baseline	0.04	0.04	0.04	0.03	0.03	0.03	0.05	0.05	0.05

Note: The number in brackets represents the size of the dataset.

References

[1] Fleuren, W., and Alkema, W., 2015, "Application of Text Mining in the Biomedical Domain," Methods, **74**(3), 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 Syst. Appl., **40**(14), pp. 5755–5764.

[3] Lloret, E., and Palomar, M., 2012, "Text Summarisation in Progress: A Literature Review," Artif. Intell. Rev., **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," J. Biomed. Inform., **52**(12), pp. 457–467.

[5] Reeve, L. H., Han, H., and Brooks, A. D., 2007, "The Use of Domain-Specific Concepts in Biomedical Text Summarization," Inf. Process. Manag., **43**(6), pp. 1765–1776.

[6] Plaza, L., Díaz, A., and Gervás, P., 2011, "A Semantic Graph-Based Approach to Biomedical Summarization," Artif. Intell. Med., **53**(1), pp. 1–14.

[7] Ji, X., Ritter, A., and Yen, P. Y., 2017, "Using Ontology-Based Semantic Similarity to Facilitate the Article Screening Process for Systematic Reviews," J. Biomed. Inform., **69**(5), pp. 33–42.

[8] Kågebäck, M., Mogren, O., Tahmasebi, N., and Dubhashi, D., 2014, "Extractive Summarization Using Continuous Vector Space Models," Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC), Gothenburg, Sweden, Apr. 26–30, pp. 31–39.

[9] Camacho-Collados, J., and Pilehvar, M. T., 2018, "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning," J. Artif. Intell. Res., **63**(1), pp. 743–788.

[10] Cheng, J., and Lapata, M., 2016, "Neural summarization by extracting sentences and words," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August.

[11] Alami, N., Meknassi, M., and En-nahnahi, N., 2019, "Enhancing Unsupervised Neural Networks-Based Text Summarization With Word Embedding and Ensemble Learning," Expert Syst. Appl., **123**(6), pp. 195–211.

[12] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., 2018, "BERT: Pre-training of deep bidirectional transformers for language understanding," North American Association for Computational Linguistics (NAACL), Minneapolis, MN, June.

[13] Zhang, H., Xu, J., and Wang, J., 2019, "Pretraining-Based Natural Language Generation for Text Summarization," In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, November, pp. 789–797.

[14] Miller, D., 2019, "Leveraging BERT for Extractive Text Summarization on Lectures," CoRR.

[15] Goldberg, Y., and Hirst, G., 2017, *Neural Network Methods for Natural Language Processing*, Vol. 10, Synth. Lect. Hum. Lang. Technol., Morgan & Claypool Publishers, pp. 1–309.

[16] Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013, "Efficient estimation of word representations in vector space," International Conference on Learning Representations: Workshops Track, Scottsdale, AZ, Jan.

[17] Ling, W., Dyer, C., Black, A. W., and Trancoso, I., 2015, "Two/too Simple Adaptations of Word2vec for Syntax Problems," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, May 31–June 5, pp. 1299–1304.

[18] Hendrycks, D., and Gimpel, K., 2016, "Bridging Nonlinearities and Stochastic Regularizers With Gaussian Error Linear Units," CoRR.

[19] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., 2018, "Deep Contextualized Word Representations," The North American Chapter of the Association for Computational Linguistics (NAACL), New Orleans, LA, June 1–6.

[20] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I., 2018, Improving Language Understanding by Generative Pre-Training, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[21] Howard, J., and Ruder, S., "Universal language model finetuning for text classification" Proceeding of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, July.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., 2017, "Attention Is All You Need," Proceeding of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, December.

[23] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., 2019, "Language Models are Unsupervised Multitask Learners," OpenAI blog, **1**(8), p. 9.

[24] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. 2020, "Language Models are Few-Shot Learners," Proceeding of 2020 Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS), Virtual, Dec. 6–12.

[25] Zhu, Q., and Luo, J., 2021, Generative Pre-Trained Transformer for Design Concept Generation: An Exploration; arXiv:2111.08489v1 [cs.CL].

[26] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S., 2015, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27.

[27] Yao, J. G., Wan, X., and Xiao, J., 2017, "Recent Advances in Document Summarization," Knowl. Inf. Syst., **53**(2), pp. 297–336.

[28] Akbik, A., Blythe, D., and Vollgraf, R., 2018, "Contextual String Embeddings for Sequence Labeling," Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, Aug. 20–26, pp. 1638–1649.

[29] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J., 2020, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," Bioinformatics, **36**(4), pp. 1234–1240.

[30] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., and McDermott, M., 2019, "Publicly available clinical BERT embeddings," Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, June.

[31] Si, Y., Wang, J., Xu, H., and Roberts, K., 2019, "Enhancing Clinical Concept Extraction With Contextual Embeddings," J. Am. Med. Inform. Assoc., **26**(11), pp. 1297–1304.

[32] Peng, Y., Yan, S., and Lu, Z., 2019, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of Bert and Elmo on Ten Benchmarking Datasets," Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, August.

[33] Beigbeder, M., and Mercier, A., 2005, "An Information Retrieval Model Using the Fuzzy Proximity Degree of Term Occurences," Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, NM, Mar. 13–17, pp. 1018–1022.

[34] Castells, P., Fernandez, M., and Vallet, D., 2006, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," IEEE Trans. Knowl. Data Eng., **19**(2), pp. 261–272.

[35] Zhang, X., Hou, X., Chen, X., and Zhuang, T., 2013, "Ontology-Based Semantic Retrieval for Engineering Domain Knowledge," Neurocomputing, **116**(9), pp. 382–391.

[36] Sanya, I. O., and Shehab, E. M., 2015, "A Framework for Developing Engineering Design Ontologies Within the Aerospace Industry," Int. J. Prod. Res., **53**(8), pp. 2383–2409.

[37] Zhang, C., Zhou, G., Lu, Q., and Chang, F., 2017, "Graph-Based Knowledge Reuse for Supporting Knowledge-Driven Decision-Making in New Product Development," Int. J. Prod. Res., **55**(23), pp. 7187–7203.

[38] Shi, F., Chen, L., Han, J., and Childs, P., 2017, "A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval," ASME J. Mech. Des., **139**(11), p. 111402.

[39] Martinez-Rodriguez, J. L., Lopez-arevalo, I., and Rios-alvarado, A. B., 2018, "OpenIE-Based Approach for Knowledge Graph Construction From Text," Expert Syst. Appl., **113**(12), pp. 339–355.

[40] Sarica, S., Luo, J., and Wood, K. L., 2020, "TechNet: Technology Semantic Network Based on Patent Data," Expert Syst. Appl., **142**(3), p. 112995.

[41] Sarica, S., and Luo, J., 2021, "Design Knowledge Representation With Technology Semantic Network," Proc. Des. Soc., **1**(7), pp. 1043–1052.

[42] Siddharth, L., Blessing, L. T. M., Wood, K. L., and Luo, J., 2021, "Engineering Knowledge Graph From Patent Database," ASME J. Comput. Inf. Sci. Eng., **22**(2), p. 021008.

[43] Hou, T., Yannou, B., Leroy, Y., and Poirson, E., 2019, "Mining Changes of User Expectations Over Time From Online Reviews," ASME J. Mech. Des., **141**(9), p. 091102.

[44] Han, J., Park, D., Forbes, H., and Schaefer, D., 2020, "A Computational Approach for Using Social Networking Platforms to Support Creative Idea Generation," Procedia CIRP, **91**, pp. 382–387.

[45] Han, Y., and Moghaddam, M., 2020, "Eliciting Attribute-Level User Needs From Online Reviews With Deep Language Models and Information Extraction," ASME J. Mech. Des., **143**(6), p. 061403.

[46] Akay, H., and Kim, S. G., 2021, "Extracting Functional Requirements From Design Documentation Using Machine Learning," Procedia CIRP, **100**, pp. 31–36.

[47] Ni, X., Samet, A., and Cavallucci, D., 2021, "Similarity-Based Approach for Inventive Design Solutions Assistance," J. Intell. Manuf., **32**(3), pp. 1–18.

[48] Gambhir, M., and Gupta, V., 2017, "Recent Automatic Text Summarization Techniques: A Survey," Artif. Intell. Rev., **47**(1), pp. 1–66.

[49] Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., and Affandy, A., 2020, "Review of Automatic Text Summarization Techniques & Methods," J. King Saud Univ. Comput. Inf. Sci., **34**(4), pp. 14–15.

[50] Mani, I., 2001, *Automatic Summarization*, Vol. 3, John Benjamins Publishing Co, pp. 221–259.

[51] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., 2002, "Bleu: A Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, July 2002.

[52] Lin, C. Y., 2004, Rouge: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out.

[53] Denkowski, M., and Lavie, A., 2014, "Meteor Universal: Language Specific Translation Evaluation for any Target Language," Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, June 26–27, pp. 376–380.

[54] Loper, E., and Bird, S., 2002, "NLTK: The Natural Language Toolkit," Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, PA, July.

[55] Kingma, D. P., and Ba, J., 2015, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), San Diego, CA, May.

[56] Reimers, N., and Gurevych, I., 2019, "Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November.
[57] Bradley, P. S., and Fayyad, U. M., 1998, "Refining Initial Points for k-Means Clustering," *ICML*, Vol. 98, pp. 91–99.
[58] Aria, H., and Vanderwende, L., 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, Boulder, CO, June.
[59] Mihalcea, R., and Tarau, P., 2004, "Textrank: Bringing Order Into Text," Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.
[60] Ozsoy, M., Alpaslan, F., and Cicekli, I., 2011, "Text Summarization Using Latent Semantic Analysis," J. Inf. Sci., **37**(4), pp. 405–417.
[61] Lin, C. Y., 2003, "Looking for a Few Good Metrics: Automatic Summarization Evaluation-how Many Samples are Enough?" Proceedings of the Fourth NII Testbeds and Community for Information Access Research Workshop (NTCIR), Tokyo, Japan, April 2003–June 2004.
[62] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K., 2017, "Text Summarization Techniques: A Brief Survey," CoRR.
[63] Gupta, V., and Lehal, G. S., 2010, "A Survey of Text Summarization Extractive Techniques," J. Emerg. Technol. Web Intell., **2**(3), pp. 258–268.
[64] Schluter, N., 2017, "The Limits of Automatic Summarisation According to Rouge," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, April.
[65] Liu, F., and Yang, L., 2008, "Correlation Between Rouge and Human Evaluation of Extractive Meeting Summaries," Proceedings of the 2008 Annual Meeting of the Association for Computational Linguistics (ACL), Columbus, OH, June.
[66] Kawin Ethayarajh, K., 2019, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November.
[67] Nguyen, Q. T., Nguyen, T. L., Luong, N. H., and Ngo, Q. H., 2020, "Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews," 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh, Vietnam, November.
[68] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., and Kurzweil, R., 2018, "Universal sentence encoder," CoRR, abs/1803.11175, Online.
[69] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A., 2017, "Supervised Learning of Universal Sentence Representations From Natural Language Inference Data," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, September.
[70] Van der Maaten, L., and Hinton, G., 2008, "Visualizing Data Using t-SNE," J. Mach. Learn. Res., **9**(11), pp. 1–27.