ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0359
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Analyzing Project Similarity in Korean Bidding Documents Using BERT

Inwoo Jung[1*], Hyunseok Moon[2], Jeongsoo Kim[3]

[1] *Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology, South Korea,* E-mail address: iwjung@kict.re.kr
[2] *Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology, South Korea,* E-mail address: hsmoon@kict.re.kr
[3] *Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology, South Korea,* E-mail address: jeongsookim@kict.re.kr

Abstract: In bidding documents, valuable information about the project exists in the form of text. When undertaking a new bid, it is necessary to refer to relevant documents from previous bidding projects, similar to the current one, to effectively understand the requirements and characteristics of the project. However, manually comparing and analyzing these documents is a time-consuming and costly process. Especially with the incorporation of emerging technologies like BIM, comparing and analyzing documents involving these new technologies requires a deeper level of expertise and understanding, posing a significant challenge. To tackle this knowledge gap, this study aims to develop a BERT-based approach to assess project similarity for Korean bidding documents. To achieve the research goal, a two-stage strategy was adopted: 1) the development of a Korean tokenizer for bidding documents in BIM technology, and 2) word embedding using BERT and project similarity analysis employing cosine similarity. The developed BERT-based similarity analysis model can automatically evaluate each project and identify the most similar project. By matching target projects with the best benchmarks, this research can assist individuals in making more accurate and timely decisions.

Key words: similarity analysis, Korean bidding documents, BERT, natural language processing

1. INTRODUCTION

A construction bidding document contains all the information necessary for bidding on a construction project, and bidders participate in the bidding of ongoing construction projects based on the bidding document. Key contents essential to be included in the bidding document comprise work details for the entire cycle of the project, such as project plan, specifications, contract conditions, bidding procedures, and bidding forms [1]. Therefore, thoroughly analyzing and accurately understanding bidding documents is a necessary process for successful bidding. There may be elements that are missed or overlooked when relying only on information about projects currently in progress. To prevent these mistakes and increase the probability of bidding success, bidders refer to past bidding records. Past bidding records allow bidders to enhance their overall strategy and plan by benchmarking others' bidding records and double-check important points when bidding on the project, therefore referring to past bidding records is as important as analyzing current bidding documents.

Recently, cutting-edge technologies are being integrated into the construction industry due to the transition to digitalization and smart construction, which has been prominent in the construction industry [2]. In particular, the scope of construction technology is gradually expanding as not only technologies in the construction domain such as Building Information Modeling (BIM) but also new

technologies in other fields such as Artificial Intelligence (AI), Internet of Things (IoT), and Digital Twin (DT) are applied to the field of construction technology [3]. This shift in the construction paradigm is also visible in the bidding documents. It is easy to see that terms and knowledge about smart construction that utilizes various cutting-edge convergence technologies are included in the documents, and bidders are required to avoid being left behind in this new trend of change. It has become necessary to increase understanding not only of traditional construction project bidding but also of smart construction project bidding. Therefore, it has traditionally been an important process for bidders to benchmark similar projects to bid on construction projects, but its importance is increasing due to the transition to smart construction.

However, it poses a significant challenge as searching through past bidding documents and identifying projects similar to the ongoing one requires a huge amount of labor [4]. In other words, a person with specialized knowledge must go through the cumbersome process of finding and analyzing all bidding documents one by one. To overcome this problem, it is necessary to develop a framework that automatically finds similar projects without specialized knowledge and allows bidders to benchmark their bids on projects currently in progress. The purpose of this study is to develop a project similarity analysis framework using Bidirectional Encoder Representations from Transformers (BERT) to help bidders effectively plan their bidding strategies.

2. BACKGROUND

2.1. Similarity Analysis in AEC domain using NLP

The approach comparing different projects is a widely used fast-track method to increase understanding of the target project by analyzing the similarities and differences between each project. This similarity analysis is essential during the bidding process, as it helps in determining the direction and specific strategies of the new bidding by referring to key requirements and considerations from the past similar project bidding records. However, since this comparison process requires significant time and cost consumptions due to its manual nature, it often faces challenges. To address this issue, Natural language processing (NLP) for document similarity analysis has been employed. The NLP allows computers to automatically find similar documents based on their text, thereby dramatically reducing the laborious task [5].

Ko [6] developed an AI-based model to evaluate project similarity using BERT and cosine similarity targeted project scope statement. The study shows that BERT model can evaluate the project similarity accurately and quickly. Erfani [7] investigated the similarity of project risks across various transportation projects using NLP and Word2vec algorithm. This study suggests that the potential for a data-driven approach to develop common risk registers while addressing project specific risks. Text similarity for railway safety and security was performed and the study highlighted that the text similarity approach can help identifying equivalent rules and procedures for safety-critical systems [8].

Through literature review, it was confirmed that NLP is being used diversely in analyzing the document similarity in the Architecture, Engineering, and Construction (AEC) domain. However, to the best of the author's knowledge, there have been few studies that have analyzed the similarity of projects based on bidding documents that reflect recent cutting-edge technologies.

2.2. Natural Language Processing: BERT

NLP is a vital area of computer science that converts human language into computational representations. With the advancement in Artificial Intelligence (AI) technology, AI-based NLP methodologies have emerged, which aims to extract not only superficial meanings but also inherent insights by analyzing natural language data such as text and speech [9].

BERT, the state-of-the-art NLP model proposed by researchers at Google AI Language in 2018, harnesses the power of bidirectional Transformer architecture to pretrain deep neural networks on extensive corpora of text, enabling it to capture contextual information and achieve exceptional performance across various NLP tasks [10]. In particular, BERT has unique pretraining approaches including masked language modeling and next sentence prediction, which provides a comprehensive understanding of contextual nuances. In addition, leveraging its pretraining on vast textual datasets,

BERT exhibits a robust understanding of language structure and semantics, allowing for efficient adaptation to diverse downstream tasks with notable performance improvements, even when training data is limited.

In contrast, research have pointed out that BERT shows the state-of-the-art performance in general NLP tasks, however, its performance deteriorates when applied to specialized domains [11]. Moreover, the most of text data used for training BERT was English-based texts, the vanilla-BERT is not suitable to deal with documents written by other languages [12].

Google AI Language has also proposed Multilingual BERT (M-BERT), which is a different version of the BERT model for multiple languages. M-BERT has an identical architecture and pretraining approaches to BERT. However, the difference from BERT is that M-BERT was trained on a corpus of 104 languages from Wikipedia. By being trained on text from numerous languages, M-BERT is able to understand text in multiple languages and has been demonstrated to be effective for various NLP tasks across a wide range of languages [13].

Considering that this study deals with BIM related bidding documents and uses Korean rather than English, the aforementioned challenges are encountered in order to carry out the project similarity analysis. To overcome these difficulties, this study employed M-BERT to develop a BERT-based framework specialized for AEC domain. In order to be able to process Korean bidding documents covering new construction technology, the developed framework learned various Korean construction technology terms, including advanced construction terms such as BIM.

3. RESEARCH METHODOLOGY

The goal of this study is to develop a BERT-based framework to evaluate project similarity of Korean bidding documents. To achieve this research objectives, a two-phase approach was adopted:

1) Developing Korean corpus and tokenizer specialized in AEC domain and
2) Developing BERT-based framework to analyze project similarity using cosine similarity.

The first phase consists of a total of 3 steps including Korean corpus collection, preprocessing, and training tokenizer based on WordPiece tokenizer. The second phase encompasses document tokenization using the trained WordPiece tokenizer, word embedding through M-BERT, and document similarity analysis using Cosine similarity. Figure 1 presents the overall process of the BERT-based framework.

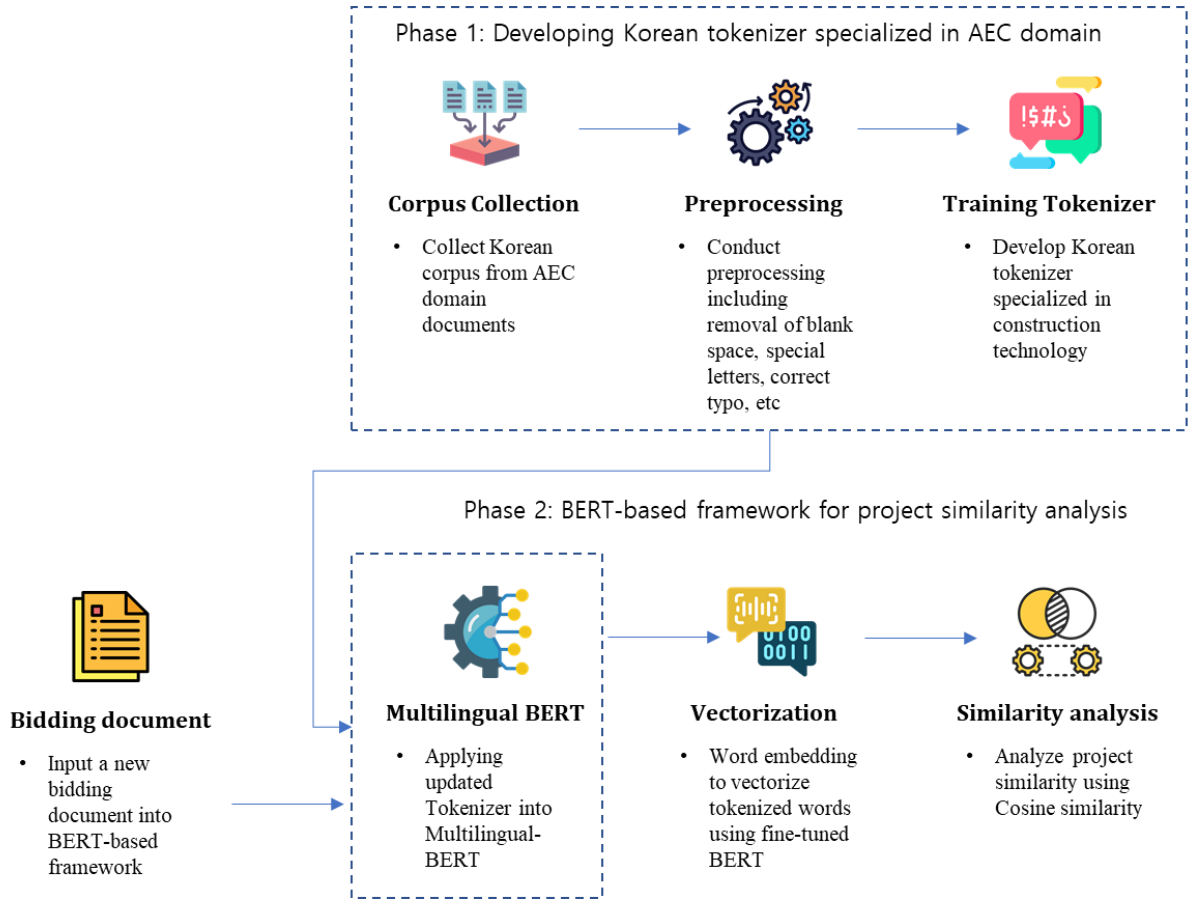

Figure 1. Overall process of BERT-based framework

3.1. Phase 1: Developing Korean tokenizer

Phase 1 begins with the collection of the Korean corpus for construction technology terms. Since the study covers not only traditional construction terms but also words related to new emerging technologies, we used multiple sources as follows: 1) BIM implementation guide for design and construction, 2) Construction dictionary, and 3) Korean Construction Standard Glossary. Additionally, an additional corpus was scraped using web-crawling technique from a dictionary of construction terms provided by the Continuous Acquisition & Life-cycle Support (CALS)[1] developed by the Korea government. In summary, a total of 20,782 construction words were collected to create a Korean corpus for training dataset. Since the created Korean corpus includes several types of noises such as blank, special letters, typos, and etc., therefore preprocessing was conducted to remove the noises using both computational method by python and manual method by authors.

The next step is to develop tokenizer specialized for the AEC domain using the preprocessed Korean corpus. M-BERT is already designed for dealing with multi-languages; however, the performance of M-BERT for other languages is not as good as BERT for English. Moreover, like BERT, M-BERT has the problem of deteriorating model performance when applied to a specific domain. To solve this problem and maximize the performance of M-BERT, we trained the tokenizer using the preprocessed Korean corpus to obtain the AEC domain-specific tokenizer. The trained tokenizer is based on WordPiece tokenizer, which is the default option for BERT. Because the WordPiece tokenizer processes words by dividing them into sub-word units, it has the advantage of being able to flexibly handle words even when rare or out-of-vocabulary words appear [14].

Figure XX shows an example of tokenization result. Original text was split into tokens using trained WordPiece tokenizer. There are special tokens that identify the beginning and end of text as “[CLS]” and “[SEP]” respectively. Although not shown in the example, the tokenizer also has other special tokens built into it, such as “[PAD]”, “[MASK]”, and “[UNK]”. “[PAD]” token is used to pad sequences to a fixed length during training and inference, and “[MASK]” token is employed in the pre-training phase for masked language modeling to indicate masked tokens. “[UNK]” token represents

[1] https://www.calspia.go.kr

unknown or out-of-vocabulary tokens in the input text. The prefix "##" attached to some tokens in the example in Figure 2 means that the token is a continuation of a word rather than the beginning of a new word. Therefore, the prefix "##" helps the M-BERT model to understand the word structure of the text and improves its performance to capture meaning and relationships between words. As a result, the developed tokenizer specializing in the construction area is used to tokenize the sentences of the bidding document before encoding them into vector values through M-BERT in the second phase.

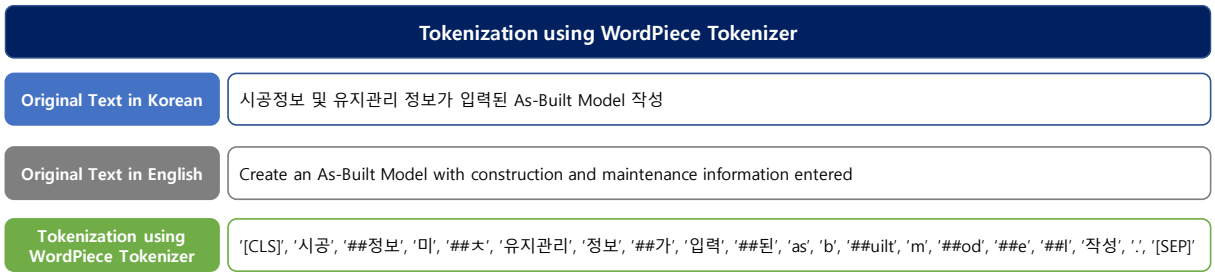

Figure 2. Example of Tokenization using WordPiece Tokenizer

3.2. Phase 2: Bert-based framework for project similarity analysis

The purpose of the Phase 2 is to extract important features from the text data and conduct a similarity analysis of documents. M-BERT was employed in this study as an encoder to vectorize Korean text data, which is not based on English language. The new bidding document is firstly preprocessed and fed into M-BERT, while the tokenizer trained in Phase 1 performs the tokenization task on the new bidding document. Subsequently, M-BERT vectorizes the tokenized text to convert natural language data into computational representations. Cosine similarity approach was used to perform project similarity analysis, and the document similarity score was calculated from the encoding values obtained through M-BERT.

Cosine similarity in NLP is a technique used to quantify the similarity between two documents by determining the cosine of the angle between the vectors of the documents [15]. For two document vectors A and B, the cosine similarity is calculated by the following formula:

$$\text{Similarity}_{\text{A,B}} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A and B are vectors of the two documents, θ is the angle between two vectors of A and B in the n-dimensional space. The range of Similarity is between 0 and 1, representing no match at all and perfect match, respectively. Therefore, the cosine similarity approach evaluates the level of similarity by calculating how closely the vectors are aligned.

4. EXPERIMENT RESULTS

This study performed similarity analysis between projects using bidding documents from five actual BIM-related projects in order to test the developed BERT-based project similarity analysis framework. Among the five projects, one project (Document 1) was assumed to be an ongoing bidding project, and the remaining four projects (Documents 2 to 5) were set as past bidding projects stored in the constructed bidding document database. The five bidding documents went through a preprocessing process to remove noise, and all text in each document was tokenized using the tokenizer developed in Phase 1. To convert the tokenized text into a form that computers can understand, word embedding was performed using the M-BERT model, and through this process, the unique features of each text were quantified. Finally, the similarity score of each document pair for Document 1 was calculated using the cosine similarity method.

Table 1 presents the results of the similarity analysis of each document with respect to Document 1. The document with the highest similarity score was document 4, recording 0.937. This means that Document 4 contains the most similar content overall to Document 1, and it will be very helpful to refer to Document 4 as well as Document 1, which is the bidding document for the ongoing project, in order to successfully carry out the ongoing bidding. Furthermore, as illustrated in Figure 3,

there is no significant difference in the similarity scores between Document 4 and Document 2, indicating that Document 2 is also highly similar to Document 1 and is therefore suitable for reference. On the other hand, Document 3 and Document 4 recorded relatively low similarity scores of 0.750 and 0.743, respectively. They can be considered relatively less important as reference documents for preparing a bid for the ongoing project.

Table 1. Results of similarity score with Document 1

Index	Title of Document	Cosine similarity score
1	Document 1*	1.000
2	Document 2	0.934
3	Document 3	0.750
4	Document 4	0.937
5	Document 5	0.743

* indicates that the bidding document is assumed to be from an ongoing project.

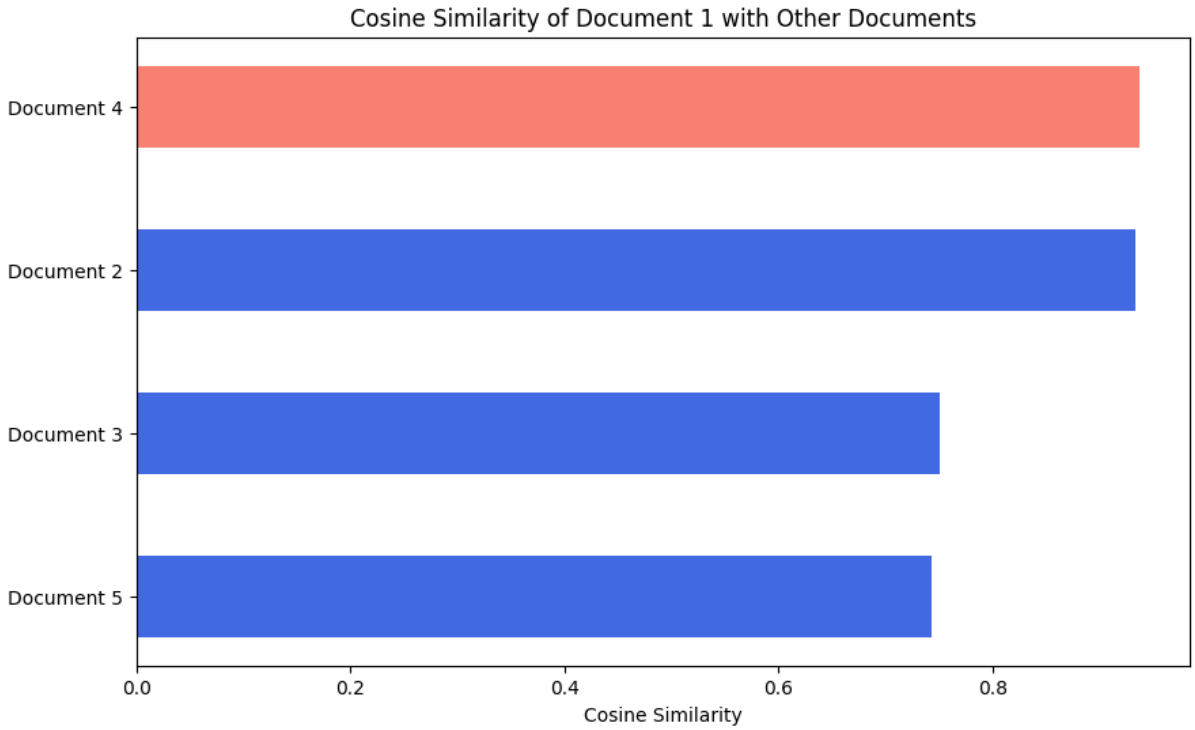

Figure 3. Similarity of Document 1 with Document database

Since similarity analysis using the cosine similarity method largely depends on the vectorization results of the text, two methods can be considered to improve the performance of the BERT-based project similarity analysis framework. First is the improvement of the tokenizer. In this study, WordPiece Tokenizer was customized by training it directly with Korean construction terms. In order to improve the performance of the tokenizer, training it with more diverse and larger amounts of language data can help improve tokenizer performance. Additionally, using a separate tokenizer specialized for Korean tokenization is a reasonable approach. Tokenization of Korean is possible with the versatile WordPiece Tokenizer, but unlike other English-based languages, Korean has a unique form and structure, so using a Korean-specific tokenizer will be effective in improving performance. The second method is to improve the encoder responsible for word embedding. M-BERT is successfully performing the task of vectorizing tokenized text, but considering the unique characteristics of the Korean language, using a model optimized for Korean language processing can improve performance of the framework. There are

various Korean-based BERT models, such as KoBERT[2] and KrBERT[3], that are based on BERT that can consider not only the meaning of the word itself but also the contextual meaning of the entire document. Therefore, additional analysis of the tokenizer and word embedding model is essential to improve the project similarity analysis framework developed in this study.

5. CONCLUSION AND FURTHER STUDY

This study proposed a BERT-based project similarity analysis framework that can identify past projects similar to the ongoing project to help bidders establish their bidding strategies. As this study targets Korean bidding documents, we used M-BERT, which can process a variety of languages, and improved the performance of the tokenizer by collecting and training the latest construction terms including cutting-edge technologies to optimize tokenization, which is one of the core processes. As a result of this framework, we were able to automatically quantify the similarity between documents, which allowed us to classify documents of high importance for reference and documents of relatively low importance. Additionally, among documents classified as important, it is possible to identify relative importance, enabling the prioritization of documents for analysis.

Although we created our own language corpus, customized the tokenizer, and adopted BERT for multiple languages to analyze Korean bidding documents, this study still has room for development. First, the language corpus needs to be expanded. A high-quality language corpus directly affects the performance of the tokenizer and has a major impact on the BERT model, which acts as an encoder. In addition, as new terms continue to appear as construction technology develops, adding newly created terms is essential for improving the quality of the project similarity analysis framework. The second is optimization of the tokenizer and word embedding method used in the framework. Since processing Korean documents, which have uniqueness distinct from English-based languages, is an important task of this research, it is necessary to maintain BERT's ability for understanding of contextual nuances while optimizing it for Korean. To this end, it is expected that by fine-tuning the existing Korean-based BERT models with a corpus of construction terms and developing them into a construction domain-specific model, it will be possible to develop a more accurate and improved project similarity analysis framework.

REFERENCES

[1] R. Shrestha, T. Ko, and J. Lee, “Uncertainties Prevailing in Construction Bid Documents and Their Impact on Project Pricing through the Analysis of Prebid Requests for Information”, Journal of Management in Engineering, 39(6), pp. 04023040, 2023.
[2] A. Pal and S.-H. Hsieh, “Deep-learning-based visual data analytics for smart construction management”, Automation in Construction, 131, pp. 103892, 2021.
[3] C. Wu, X. Li, Y. Guo, J. Wang, Z. Ren, M. Wang, and Z. Yang, “Natural language processing for smart construction: Current status and future directions”, Automation in Construction, 134, pp. 104059, 2022.
[4] N. Torkanfar and E.R. Azar, “Quantitative similarity assessment of construction projects using WBS-based metrics”, Advanced Engineering Informatics, 46, pp. 101179, 2020.
[5] Y. Zou, A. Kiviniemi, and S.W. Jones, “Retrieving similar cases for construction project risk management using Natural Language Processing techniques”, Automation in construction, 80, pp. 66-76, 2017.
[6] T. Ko, H. David Jeong, and J. Lee, “Natural language processing–driven similar project determination using project scope statements”, Journal of Management in Engineering, 39(3), pp. 04023005, 2023.
[7] A. Erfani, Q. Cui, and I. Cavanaugh, “An empirical analysis of risk similarity among major transportation projects using natural language processing”, Journal of Construction Engineering and Management, 147(12), pp. 04021175, 2021.

[2] https://github.com/SKTBrain/KoBERT

[3] https://github.com/snunlp/KrBERT

[8] A.W. Qurashi, V. Holmes, and A.P. Johnson. “Document processing: Methods for semantic text similarity analysis”. in 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2020.
[9] M. Vierlboeck, D. Dunbar, and R. Nilchiani. “Natural Language Processing to Extract Contextual Structure from Requirements”. in 2022 IEEE International Systems Conference (SysCon), 2022.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.
[11] M. Ryu, G. Lee, and K. Lee, “Knowledge distillation for bert unsupervised domain adaptation”, Knowledge and Information Systems, 64(11), pp. 3113-3128, 2022.
[12] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, “Kr-bert: A small-scale korean-specific language model”, arXiv preprint arXiv:2008.03979, 2020.
[13] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?”, arXiv preprint arXiv:1906.01502, 2019.
[14] C. Toraman, E.H. Yilmaz, F. Şahinuç, and O. Ozcelik, “Impact of tokenization on language models: An analysis for turkish”, ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), pp. 1-21, 2023.
[15] W. Chang, Z. Xu, S. Zhou, and W. Cao, “Research on detection methods based on Doc2vec abnormal comments”, Future Generation Computer Systems, 86, pp. 656-662, 2018.