IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 January 2023, accepted 23 February 2023, date of publication 2 March 2023, date of current version 13 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251664

RESEARCH ARTICLE

Examine the Effectiveness of Patent Embedding-Based Company Comparison Method

TAEHYUN HA[1,2] **AND JAE-MIN LEE**[2]
[1]Department of Data Science, Sejong University, Seoul 02456, Republic of Korea
[2]Future Technology Analysis Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea

Corresponding author: Jae-Min Lee (jmlee@kisti.re.kr)

This work was supported by the Korea Research Institute of Science and Technology Information under Grant K-23-L03-C02-S01.

ABSTRACT A company's benchmarking strategy is significantly determined by how it measures technological similarity. Researchers have measured the technological similarity between companies using a vector composed of the classification codes of patents that each company owns. However, patent classification code-based company comparison methods do not consider the text in patents and thus may not find similar companies accurately. To solve this problem, this study suggests a patent embedding-based company comparison method. The suggested method uses a text embedding model to vectorize the text in patents and calculates technological similarity based on the embedding vector. We examine the effectiveness of the suggested method by comparing it with the conventional patent classification code-based method. From the validation results for 11,227 Korean companies listed in the Korea Data Analysis, Retrieval, and Transfer system (DART), we find that the suggested method effectively retrieves technologically similar companies.

INDEX TERMS Patents, R&D benchmarking, text embedding, technological similarity.

I. INTRODUCTION

The level of technology plays an important role in determining the competitiveness of a company. Among the various items that represent the research and development (R&D) performance of a company, previous studies have mainly focused on patents. Patents legally protect the right to technology use and help to occupy an exclusive position of a company. Thus, companies have tried to enroll patents by describing their novel technologies and specifying their rights based on the patents. People have utilized patents to measure the R&D performances of companies and to analyze the technological competitors of a company. Based on the precise identification of the R&D performance and the technological competitors of a company, stakeholders can consider an effective strategic approach to improve their technological competitiveness.

Focusing on the importance of patents, previous studies have suggested various methods for analyzing patents. Roughly, these studies can be classified as relation-based and attribute-based approaches. The relation-based approach mainly focuses on the citations of patents. Because this approach does not require the attributes of patents, patents with insufficient descriptions can be analyzed if their citations are identified. While analyzing patent citations, researchers identified significant changes in certain technology fields [1]. However, this approach was not effective for the latest patents because the patents did not have an adequate number of citations to represent their characteristics. Contrary to the relation-based approach, the attribute-based approach focuses on the attributes of patents such as the title, abstract, claim, and classification codes. Because this approach does not rely on citations, the latest patents can be analyzed by focusing on their content. Previous studies have shown that the attributes of patents can be utilized to discover prominent patents and emerging technologies (e.g., [2], [3]). However, due to the lack of precise methods for analyzing the description of patents, their performances were limited to a certain level.

The limitations of the attribute-based approach have been overcome by a notable advance in text embedding techniques. Mikolov, Chen, Corrado, and Dean [4] suggested an effective technique for representing the semantics of words in the vector space named Word2Vec, and Le and Mikolov [5] extended the word-level approach to a sentence-level approach named

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Jin.

Doc2Vec. Later studies considered a more effective way to train the embedding model. Peters et al. [6] found that the embedding model could be trained more effectively when the model was initially trained on general texts (e.g., news articles) and retrained on task-oriented texts. This embedding model was named Embedding from the Language MOdel (ELMO), and its pretraining and fine-tuning approach has become a major text embedding technique. Devlin et al. [7] employed the pretraining and fine-tuning approach and suggested a more precise embedding model named Bidirectional Encoder Representations from Transformers (BERT). BERT has facilitated the analysis of complex texts and has been utilized in various fields including chemistry [8], biomedicine [9], and general science [10]. However, only a few studies employed the technique to analyze the R&D performance of companies in terms of patents.

The text embedding technique can find technologically similar companies accurately and can be used to develop an effective R&D benchmarking service. Our study developed a patent embedding model and examined its effectiveness for retrieving technologically similar companies. We designed BERT models with different settings and trained them on the descriptions (i.e., titles, abstracts, and claims) and classification codes of Korean patents. The model that showed the best performance was selected and used to develop a patent embedding-based company comparison method. For 11,227 Korean companies listed in the Korea Data Analysis, Retrieval, and Transfer system (DART), we compared the performance of the suggested patent embedding-based company comparison method and conventional patent classification code-based company comparison method, and we present the effectiveness of the suggested method.

II. BACKGROUND

A. PATENT ANALYSIS

Patents contain descriptions of inventions. The forms of patents are somewhat different by country, but in general, patents are described by their title, abstract, and claims, and their classification codes follow the International Patent Classification (IPC) or Cooperative Patent Classification (CPC) system. Like other general documents, the titles and abstracts of patents describe their overall contents. However, unlike other documents, patents have a unique entry named claims. Claims specify the idea that patentees want to claim their rights by the patent. Claims consist of independent and dependent claims that describe the points of the claims, generally and in detail, respectively. To facilitate the identification of patents, the IPC system was established by the Strasbourg Agreement in 1971. This system is composed of approximately 70 thousand codes and is renewed annually. However, because the system is limited to covering a broad range of technology, the United States Patent and Trademark Office (USPTO) and European Patent Office (EPO) developed a new classification system in 2012 named CPC. The CPC system provides approximately 250 thousand codes with a hierarchy architecture that consists of sections, classes, subclasses, groups, and subgroups.

Previous studies have used patent classification codes to find technology competitors and alternatives. Jaffe [11] suggested that the distribution of patent classification codes can be used to find similar companies in terms of technology. He calculated cosine similarity for the distributions of patent classification codes and defined it as the technological proximity between two companies. Kay et al. [12] used IPC codes to build patent overlay maps between two companies and showed that the maps were effective in benchmarking. Lee and Lee [13] also employed IPC codes to measure the technology similarity of companies and used them to draw a patent analysis map for the purpose of benchmarking. As the CPC system was launched to replace the IPC system, several studies have suggested that CPC codes are a better tool with which measure technological similarities among companies. Kapoor et al. [14] examined patent portfolios in the wind industry and found that the CPC system was more effective in capturing the technological features of wind power companies than was the IPC system. Leydesdorff et al. [15] also demonstrated the effectiveness of CPC codes for capturing similar patent portfolios. Kim and Bae [16] used CPC codes to measure the similarity of patents and constructed technology clusters to capture emerging technologies. Jee et al. [17] employed CPC codes to find different types of technology clusters among patents.

On the other hand, other researchers have tried to reflect the contents of patents, such as titles, abstracts, and claims, to retrieve similar patents. For example, some studies used Term Frequency and Inversed Document Frequency (TF-IDF) to find emerging technologies (e.g., [18], [19]). Singular Value Decomposition (SVD, [20]), Latent Dirichlet Allocation (LDA, [21]), Support Vector Machine (SVM, [22]), and Artificial Neural Network (ANN, [23]) have been employed to find technological trends and opportunities. However, because of the high complexity in the text of patents, these approaches have limited accuracy. Other studies have tried to overcome the limitation by considering additional features such as numbers of claims, citations, classification codes, and inventors of patents (e.g., [2], [3]). However, as the number of features increases, some of the features may not be available for some patents, and overfitting problems can occur.

B. TEXT EMBEDDING

Recent advances in text embedding techniques have notably increased the accuracy of Natural Language Processing (NLP). With the use of the text embedding technique, unique characteristics of patents can be captured even if the patents are recent. Early text embedding techniques were only able to consider word-level semantics, but later techniques were able to consider sentence-level semantics. Several techniques, including Word2Vec [4], Doc2Vec [5], and ELMO [6], have been suggested for achieving high accuracies on NLP tasks. In particular, BERT [7] has shown state-of-the-art

performances on several NLP tasks and has been employed as a base model for later extended models. Based on a pretrained BERT model, several researchers have suggested fine-tuned BERT models with different fields of research in chemistry [8], biomedicine [9], and general science [10].

In the context of patent analysis, several studies have utilized text embedding techniques to extract feature vectors from the patent text. Li et al. [24] designed a Convolutional Neural Network (CNN) model to embed the text of patents in the USPTO dataset and showed that the model performed better than naïve Bayes, random forest, decision tree, and simple ANN methods. Chen et al. [25] used topic modeling and word embedding techniques to develop a patent recommendation system and reported that the model performed better than did the Latent Semantic Analysis (LSA), Word2Vec, and Doc2Vec models. Some studies examined the patents of other countries. For example, Kim et al. [26] used a Korean patent dataset to examine a novel clustering method based on Doc2Vec and showed that the method performed better than clustering based on the TF-IDF, K-means, and Doc2Vec methods. Zhu et al. [27] considered a Chinese patent dataset to examine the performance of a CNN-based patent classification model and reported superior performances of the model compared with the Recurrent Neural Network (RNN), SVM, Bayesian, and regression models.

However, despite the cases, only a few cases employed a BERT model to compare companies in terms of patents and then developed an R&D benchmarking service based on the model. Lee and Hsiang [28] examined the performance of BERT with USPTO patent datasets and showed that training BERT on patent claims and CPC codes achieved the best performance compared with cases using titles, abstracts, and IPC codes. Kang et al. [29] tested BERT on a patent dataset for dual-camera technology and found that BERT could be utilized for searching patent prior art. These studies demonstrated the potential of BERT as a useful tool for patent analysis. However, none of them considered the use of BERT for finding similar companies in terms of patents. A company's patents can be vectorized using BERT and utilized for comparison purposes. This can be especially useful for finding certain companies to benchmark and identify technological competitiveness.

III. METHODS

We designed a stepwise approach to develop a patent embedding-based company comparison method, and we validated its effectiveness. At first, we examined BERT models with different settings, and we selected the best one. Then, we devised a patent embedding-based company comparison method that extracts patent embedding vectors and calculates technological similarity based on the embedding vectors. Lastly, the devised method was compared with the conventional patent classification code-based company comparison method in terms of its effectiveness in retrieving technologically similar companies. Figure 1 shows the stepwise approach in detail.

A. DATA COLLECTION AND PREPROCESSING

Korea Institute of Patent Information (KIPI) provides a Korean patent database named Korea Intellectual Property Rights Information Service (KIPRIS). We secured 1,533,915 patents that were registered by 2020 for 32,209 companies listed with the KOrea Industrial Technology Association (KOITA). The titles, abstracts, and first claims of patents were considered as the input sources of the BERT models (although patents have multiple independent and dependent claims, we only considered the first claim because of the simplicity issue [28]). We cleaned texts in the dataset using the Natural Language ToolKit (NLTK) package and we checked for any misused expressions and excluded them from the dataset. For the supervised learning of BERT models, we considered the CPC classification codes of patents as output sources. Because the CPC codes consisted of approximately 250 thousand codes, we only considered the subclass level CPC codes (four digits). Similar to the preprocessing of patent texts, we manually checked the CPC codes of patents and excluded misused expressions. Indexing scheme codes were also removed from the dataset because they were not functional in terms of patent classification. Table 1 shows the basic statistics of the dataset.

B. MODEL TRAINING AND VALIDATION

Google research provides a basic BERT model pretrained on English texts and other BERT models pretrained on multilingual texts. However, these models were not generally oriented to Korean. We employed a different BERT model named KoBERT. KoBERT was pretrained on 5 million sentences and 54 million words from Wikipedia in Korean and released by SKT Brain (https://github.com/SKTBrain/KoBERT). KoBERT has been employed in various studies for embedding texts in different contexts (e.g., [30], [31]). We trained seven KoBERT models on different sets of texts consisting of (1) titles, (2) abstracts, (3) claims, (4) titles+abstracts, (5) titles+claims, (6) abstracts+claims, and (7) titles+abstracts+claims as input sources, and CPC codes as an output source. The parameters of the KoBERT models, which consisted of maximum sequence length, batch size, and learning rate, were set to 128, 256, and 2e-5, respectively.

To conduct fine-tuning with KoBERT using supervised learning, a proper fully connected layer and loss function had to be applied to the basic KoBERT model. We added a 768×651 fully connected layer to the end of the basic KoBERT model. Because each patent can have multiple CPC codes, we regarded the fine-tuning task as a multilabel classification problem, and we used binary cross-entropy and sigmoid functions to calculate losses in the training and test processes. The Adam optimizer was employed for training the models. We divided the dataset into 70% data for training and 30% for testing and used them to measure performance. From among the trained KoBERT models with different settings, we selected the best one in terms of F1 and accuracy score.

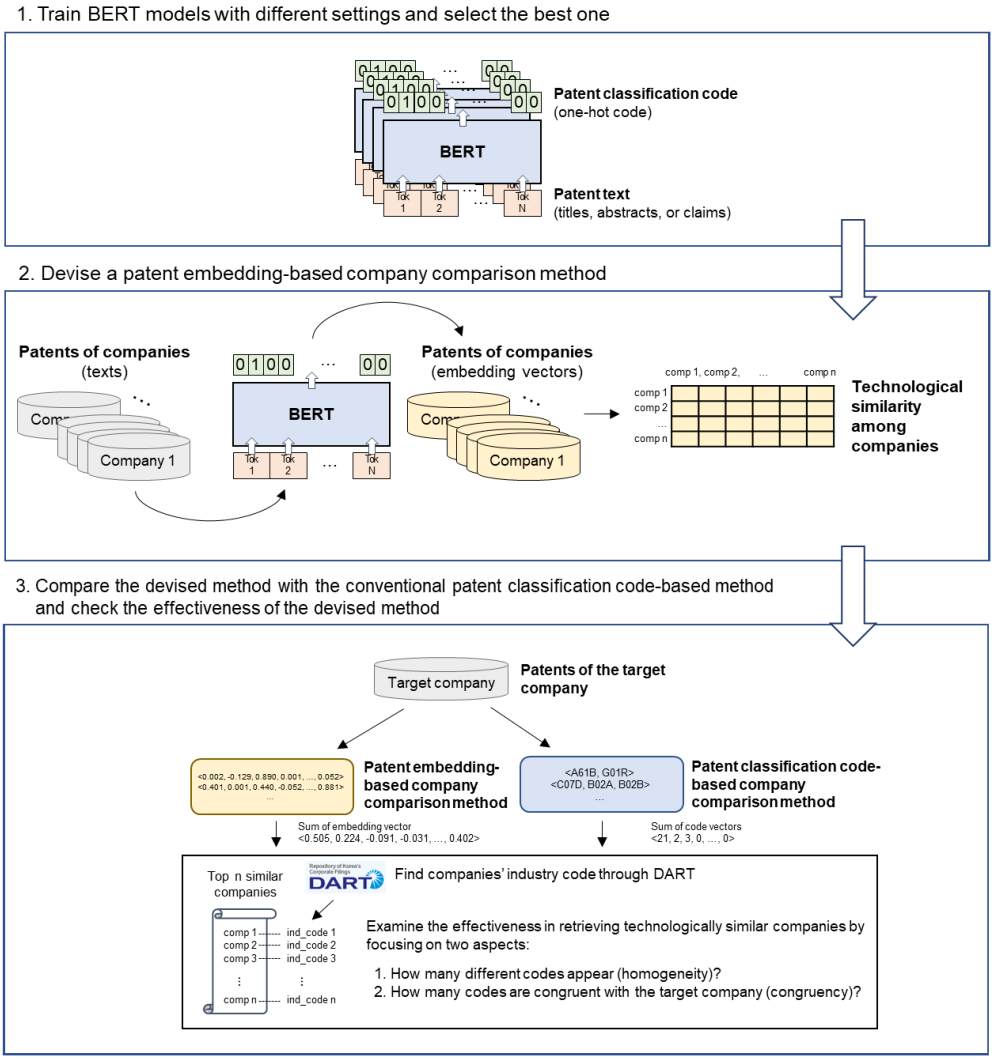

FIGURE 1. Method development and validation process.

TABLE 1. Basic statistics of the dataset.

	Number of unique values	Mean of the text length	Std. dv. of the text length	Minimum text length	Maximum text length
Title	2,080,230	21.250	13.229	2	348
Abstracts	2,493,359	373.650	632.359	3	1,253,263
First claims	2,513,046	542.758	799.114	2	261,451
CPC codes	651	4	0	4	4

* Note: Std. dv. indicates standard deviation.

Accuracy was calculated based on cases of correctly predicted (i.e., all CPC codes of patents are correctly predicted) or not (i.e., at least one CPC code is wrongly predicted). Thus, the accuracy score was expected to be lower than that of normal classification problems.

To check whether the model learned the patent information correctly, we evaluated its effectiveness in finding technologically similar companies. Previous studies have employed patent classification code-based methods to find technologically similar companies (e.g., [11], [12], [13]). The methods identify companies' patents and their classification codes to construct code vectors and sum vectors for measuring technological similarity between companies. However, the methods have resolution and sparsity problems. Depending on the level of classification codes, the vector dimension can be too high or too low (e.g., the number of CPC codes is approximately 250 thousand). Also, if a company's patents were focused on specific technology fields,

TABLE 2. Model performances.

	Training data				Test data			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
Title	0.394	0.619	0.632	0.606	0.338	0.503	0.472	0.538
Abstract	0.452	0.670	0.670	0.671	0.387	0.547	0.507	0.595
Claim	0.226	0.376	0.373	0.380	0.218	0.331	0.310	0.355
Title+Abstract	0.452	0.668	0.668	0.669	0.397	0.556	0.517	0.602
Title+Claim	0.464	0.683	0.683	0.682	0.381	0.540	0.495	0.594
Abstract+Claim	0.465	0.685	0.684	0.686	0.374	0.534	0.490	0.589
Title+Abstract+Claim	0.447	0.665	0.664	0.665	0.385	0.545	0.507	0.590

only a few classification codes would appear, and this can cause the sparsity problem. We considered that the trained KoBERT model with patent information could address the problems effectively, and we devised the patent embedding-based company comparison method.

The patent embedding-based company comparison method extracts embedding vectors for companies' patents (e.g., 761-dimensional vectors in the case of our trained KoBERT model) and uses sum vectors to calculate the technological similarity between companies (cosine similarity). We compared the effectiveness of the method with that of the CPC-based company comparison method. In detail, we used 11,227 companies listed in the KOITA and DART databases and identified their industry codes using DART. DART is a system that is managed by the Financial Supervisory Service (FSS) in Korea, and it provides detailed information about companies such as name, stock code, representative name, corporate and business registration codes, and industry code.

With the patent embedding-based and CPC-based methods, we checked the top 100 and 10 results for similar company retrieval, and we examined how many different industry codes appeared in a result (homogeneity) and how many codes in a result were congruent with the target company (congruency). If the technological features of companies in a retrieval result were homogeneous and congruent with the target company, only a few industry codes would appear, and many industry codes would be congruent with the target company. In contrast, if the technological features were heterogeneous and incongruent with the target company, various industry codes would appear, and only a few would be congruent with the target company.

The company industry codes are based on the Korea Standard Industry Code (KSIC). KSIC is composed of four-level, five-digit codes. The first two numbers of the code indicate the first-level classification, and the third, fourth, and fifth numbers of the code indicate the second-, third-, and fourth-level classifications, respectively. For example, industry code 25931 represents that the company is hierarchically classified into 25 (manufacture of fabricated metal products, except machinery and furniture), 259 (manufacture of other fabricated metal products; metalworking service activities), 2593 (manufacture of cutlery, hand tools and general hardware) and 25931 (manufacture of cutlery). Considering the hierarchical structure of industry codes, we examined the homogeneity and congruency scores at different levels of industry codes.

IV. RESULTS

Among the seven models examined, we found that the KoBERT model fine-tuned with abstracts and claims performed the best in training data and the model fine-tuned with titles and abstracts performed the best in test data. We considered that the latter model showed better generalization performance, and we used it to develop a patent embedding-based company comparison method. Table 2 shows the result. The effectiveness of the developed method was compared with the CPC-based company comparison method, and Table 3 and Table 4 show the results. As shown in the table, the patent embedding-based method presented lower means for the number of unique industry codes and higher means for the number of congruent industry codes than did the CPC-based method at the two-, three-, four-, and five-digit industry code levels. This tendency was presented consistently for all levels of industry codes and top 100 and 10 results. This implies that the patent embedding-based method found technologically similar companies more effectively than did the CPC-based method. The industry codes retrieved by the patent-embedding method were less dispersed but more congruent with the target company than the codes of the CPC-based method.

V. DISCUSSION

We considered KoBERT with different settings to find a proper patent embedding model, and the results showed that title+abstract was the best source of patent information for training the KoBERT model. Our model showed accuracies of 0.452 and 0.397, and F1 scores of 0.668 and 0.556 on the training and test data, respectively. Lee and Hsiang [28] reported that a BERT model fine-tuned with claims and four-digit CPC codes of patents achieved an F1 score of 0.668, which is very close to the performance of our model. Considering that the previous study mainly used English patents to train the model, this implies that the language issue may not present a major problem in training the patent embedding model. We were not able to fully compare our model evaluation results with those of the previous study because the previous studies did not examine the effect of patent text type (they only focused on claims rather than titles and abstracts).

TABLE 3. Comparison of CPC and patent embedding-based methods (top 100).

	Two-digit code (First level)		Three-digit code (Second level)	
	Embedding-based	CPC-based	Embedding-based	CPC-based
Mean number of unique industry codes (Homogeneity)	13.929	15.293	23.545	25.144
Mean number of congruent industry codes (Congruency)	29.944	26.809	19.259	17.180
	Four-digit code (Third level)		Five-digit code (Fourth level)	
	Embedding-based	CPC-based	Embedding-based	CPC-based
Mean number of unique industry codes (Homogeneity)	34.834	36.440	45.082	46.726
Mean number of congruent industry codes (Congruency)	12.305	11.030	8.120	7.375

TABLE 4. Comparison of CPC and patent embedding-based methods (top 10).

	Two-digit code (First level)		Three-digit code (Second level)	
	Embedding-based	CPC-based	Embedding-based	CPC-based
Mean number of unique industry codes (Homogeneity)	4.005	4.175	5.129	5.264
Mean number of congruent industry codes (Congruency)	3.885	3.459	2.748	2.406
	Four-digit code (Third level)		Five-digit code (Fourth level)	
	Embedding-based	CPC-based	Embedding-based	CPC-based
Mean number of unique industry codes (Homogeneity)	6.174	6.260	7.006	7.113
Mean number of congruent industry codes (Congruency)	1.973	1.706	1.388	1.184

However, we found only small differences in the performance of models that were fine-tuned on different text types. This suggests that not the patent text type, but other issues in the data, such as the quantity of data and the coverage, could be more important in training a patent embedding model.

The effectiveness of patent embedding models has been investigated in several studies. These studies have reported that text embedding models can be used effectively to classify patents into certain patent classification codes ([24], [28]). However, these studies did not explain how the models could be utilized to retrieve technologically similar companies and how effective this method is compared with existing methods. To address this issue, we devised a patent embedding-based company comparison method and compared its effectiveness with the conventional patent classification code-based company comparison method. The effectiveness of the methods was measured in terms of homogeneity (how many different industry codes of companies appeared in the retrieval results of the target company) and congruency (how many companies had the same industry code as the target company). The results showed that the devised patent embedding-based method performed better than the conventional one. However, we need to note that this validation could be limited because only a single industry code was assigned to a company even if the company had a broad technology portfolio. For example, industry code 26400 (telecommunication and broadcast equipment manufacturing) is assigned to Samsung Electronics, but other industry codes such as 26100 (semiconductor manufacturing) and 46520 (wholesale business of home appliance and telecommunication equipment) could be assigned too. In other words, the validity of the method can be perceived differently depending on what people judge to be the company's main technology area.

We also note that the properties of patents may be different depending on how patentees describe their ideas and claims in the patents. If they describe key ideas using figures, the patent texts provide less important information. In this case, the patent classification code-based method could be more effective than the patent embedding-based method. One could consider a hybrid approach to cover various types of patents. The approach constructs a vector combining patent embedding and classification code vectors and utilizes it to find technologically similar companies. The number of figures or the length of the title, abstract, and claims in the patents can be utilized to determine a proper ratio that combines the patent embedding and classification code vectors. In addition, researchers can consider different models for embedding patent texts and examine their classification performance and effectiveness in retrieving technologically similar companies. Extended patent data that contains more data for a broader range of patents would help improve performance and effectiveness.

VI. CONCLUSION

In this study, we examined the effectiveness of a patent embedding-based method for finding technologically similar companies. To devise the patent embedding-based method, we examined KoBERT models that had different settings and selected the best model. The examination results showed that title+abstract was the best source for training the

patent embedding model. The devised patent embedding-based method was compared with the conventional patent classification code-based company comparison method. The validation results demonstrated that the devised method more effectively found technologically similar companies than the conventional method. Although we found that title+abstract was the best source for training the model, we observed only small differences between the performances of the examined models. This implies that the type of patent texts did not significantly impact the performance. We suggest that other properties of patent data, such as size and coverage, could be more important for determining performance. In addition, advances in embedding models and techniques would contribute to establish the patent embedding-based method successfully. Future studies can address the issues and suggest an improved approach.

REFERENCES

[1] P. Sharma and R. C. Tripathi, "Patent citation: A technique for measuring the knowledge flow of information and innovation," *World Pat. Inf.*, vol. 51, pp. 31–42, Dec. 2017.

[2] M. N. Kyebambe, G. Cheng, Y. Huang, C. He, and Z. Zhang, "Forecasting emerging technologies: A supervised learning approach through patent analysis," *Technol. Forecasting Social Change*, vol. 125, pp. 236–244, Dec. 2017.

[3] C. Lee, O. Kwon, M. Kim, and D. Kwon, "Early identification of emerging technologies: A machine learning approach using multiple patent indicators," *Technol. Forecasting Social Change*, vol. 127, pp. 291–303, Feb. 2018.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[5] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, Jun. 2014, pp. 1–9.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[8] J. Payne, M. Srouji, D. A. Yap, and V. Kosaraju, "BERT learns (and teaches) chemistry," 2020, *arXiv:2007.16012*.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[10] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.

[11] A. B. Jaffe, "Characterizing the 'technological position' of firms, with application to quantifying technological opportunity and research spillovers," *Res. Policy*, vol. 18, no. 2, pp. 87–97, Apr. 1989.

[12] L. Kay, N. Newman, J. Youtie, A. L. Porter, and I. Rafols, "Patent overlay mapping: Visualizing technological distance," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 12, pp. 2432–2443, Dec. 2014.

[13] M. Lee and S. Lee, "Identifying new business opportunities from competitor intelligence: An integrated use of patent and trademark databases," *Technol. Forecasting Social Change*, vol. 119, pp. 170–183, Jun. 2017.

[14] R. Kapoor, M. Karvonen, S. Ranaei, and T. Kässi, "Patent portfolios of European wind industry: New insights using citation categories," *World Pat. Inf.*, vol. 41, pp. 4–10, Jun. 2015.

[15] L. Leydesdorff, D. F. Kogler, and B. Yan, "Mapping patent classifications: Portfolio and statistical analysis, and the comparison of strengths and weaknesses," *Scientometrics*, vol. 112, no. 3, pp. 1573–1591, Jul. 2017.

[16] G. Kim and J. Bae, "A novel approach to forecast promising technology through patent analysis," *Technol. Forecasting Social Change*, vol. 117, pp. 228–237, Apr. 2017.

[17] J. Jee, H. Shin, C. Kim, and S. Lee, "Six different approaches to defining and identifying promising technology through patent analysis," *Technol. Anal. Strategic Manage.*, vol. 34, no. 8, pp. 961–973, Aug. 2022.

[18] J. Joung and K. Kim, "Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data," *Technol. Forecasting Social Change*, vol. 114, pp. 281–292, Jan. 2017.

[19] H. Niemann, M. G. Moehrle, and J. Frischkorn, "Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application," *Technol. Forecasting Social Change*, vol. 115, pp. 210–220, Feb. 2017.

[20] D. S. Kwon, D. Li, and S. Y. Sohn, "Identifying innovation in socialist countries through patent analysis focused on Cuba and Vietnam," *World Pat. Inf.*, vol. 59, Dec. 2019, Art. no. 101898.

[21] D. Choi and B. Song, "Exploring technological trends in logistics: Topic modeling-based patent analysis," *Sustainability*, vol. 10, no. 8, p. 2810, Aug. 2018.

[22] J. Yun and Y. Geum, "Automated classification of patents: A topic modeling approach," *Comput. Ind. Eng.*, vol. 147, Sep. 2020, Art. no. 106636.

[23] M. H. Ramadhan, V. I. Malik, and T. Sjafrizal, "Artificial neural network approach for technology life cycle construction on patent data," in *Proc. 5th Int. Conf. Ind. Eng. Appl. (ICIEA)*, Singapore, Apr. 2018, pp. 499–503.

[24] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: Patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, Sep. 2018.

[25] J. Chen, J. Chen, S. Zhao, Y. Zhang, and J. Tang, "Exploiting word embedding for heterogeneous topic model towards patent recommendation," *Scientometrics*, vol. 125, no. 3, pp. 2091–2108, Aug. 2020.

[26] J. Kim, J. Yoon, E. Park, and S. Choi, "Patent document clustering with deep embeddings," *Scientometrics*, vol. 123, no. 2, pp. 563–577, Mar. 2020.

[27] H. Zhu, C. He, Y. Fang, B. Ge, M. Xing, and W. Xiao, "Patent automatic classification based on symmetric hierarchical convolution neural network," *Symmetry*, vol. 12, no. 2, p. 186, Jan. 2020.

[28] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning BERT language model," *World Pat. Inf.*, vol. 61, Jun. 2020, Art. no. 101965.

[29] D. M. Kang, C. C. Lee, S. Lee, and W. Lee, "Patent prior art search using deep learning language model," in *Proc. 24th Symp. Int. Database Eng. Appl.*, Seoul, South Korea, Aug. 2020, pp. 1–5.

[30] E. Kim, H. Yoon, J. Lee, and M. Kim, "Accurate and prompt answering framework based on customer reviews and question-answer pairs," *Expert Syst. Appl.*, vol. 203, Oct. 2022, Art. no. 117405.

[31] H. Kim, J. Namgung, S. Son, M.-S. Gil, and Y.-S. Moon, "Performance comparison of spoken language detection models with embedding replacement," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2021, pp. 106–109.