ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0375
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Application of Domain-specific Thesaurus to Construction Documents based on Flow Margin of Semantic Similarity

Youmin PARK[1], Seonghyeon MOON[2,*], Jinwoo KIM[3], Seokho CHI[4]

[1] *Department of Industrial and Systems Engineering, Gyeongsang National University, South Korea,* E-mail address: ympark1106@naver.com
[2] *Corresponding Author; Department of Industrial and Systems Engineering, Gyeongsang National University, South Korea,* E-mail address: moonsh@gnu.ac.kr
[3] *Department of Civil & Environmental Engineering, Hanyang University, South Korea,* E-mail address: jinwookim@hanyang.ac.kr
[4] *Department of Civil and Environmental Engineering, Seoul National University, Seoul, South Korea,* E-mail address: shchi@snu.ac.kr

Abstract: Large Language Models (LLMs) still encounter challenges in comprehending domain-specific expressions within construction documents. Analogous to humans acquiring unfamiliar expressions from dictionaries, language models could assimilate domain-specific expressions through the use of a thesaurus. Numerous prior studies have developed construction thesauri; however, a practical issue arises in effectively leveraging these resources for instructing language models. Given that the thesaurus primarily outlines relationships between terms without indicating their relative importance, language models may struggle in discerning which terms to retain or replace. This research aims to establish a robust framework for guiding language models using the information from the thesaurus. For instance, a term would be associated with a list of similar terms while also being included in the lists of other related terms. The relative significance among terms could be ascertained by employing similarity scores normalized according to relevance ranks. Consequently, a term exhibiting a positive margin of normalized similarity scores (termed a pivot term) could semantically replace other related terms, thereby enabling LLMs to comprehend domain-specific terms through these pivotal terms. The outcome of this research presents a practical methodology for utilizing domain-specific thesauri to train LLMs and analyze construction documents. Ongoing evaluation involves validating the accuracy of the thesaurus-applied LLM (e.g., S-BERT) in identifying similarities within construction specification provisions. This outcome holds potential for the construction industry by enhancing LLMs' understanding of construction documents and subsequently improving text mining performance and project management efficiency.

Key words: Domain-specific Thesaurus; Construction Document; Semantic Similarity; Natural Language Processing; Large Language Model

1. INTRODUCTION

The construction industry is fundamentally experiential, with onsite experiential knowledge being documented for preservation. The analysis of such experiential knowledge has led to a surge in text mining research aimed at extracting valuable insights from documented data. With the advent of Large Language Models (LLMs), there has been significant progress in text mining research, prompting attempts to apply LLMs within the construction industry as well.

However, a major challenge has tackled LLMs' application in understanding the specialized terminology and expressions in the construction domain. To address this issue, numerous studies have explored the use of thesauri, dictionaries that define relationships between words, to forcibly teach domain-specific expressions to language models. Although thesauri offer the advantage of acquainting language models with intricate expressions, practical limitations still exist. Thesauri merely inform about the relationships between words without specifying their relative importance, leaving models at a loss on which words to retain or substitute with thesaurus terms during text data comprehension.

This research aims to present a concrete method for applying thesauri in the text mining of construction documents using language models. Our process involves several steps. Initially, we construct a thesaurus using the Word2Vec model, known for its superior performance in analyzing word relationships. Subsequently, we determine the relative importance of words based on the PageRank algorithm. Then, employing the Flow Margin algorithm, we guide the language model on which words to keep and which to replace with thesaurus terms during text data analysis(Figure 1). Finally, we visualize the relationships between words (which to retain and which to substitute) in the form of a Word Network.

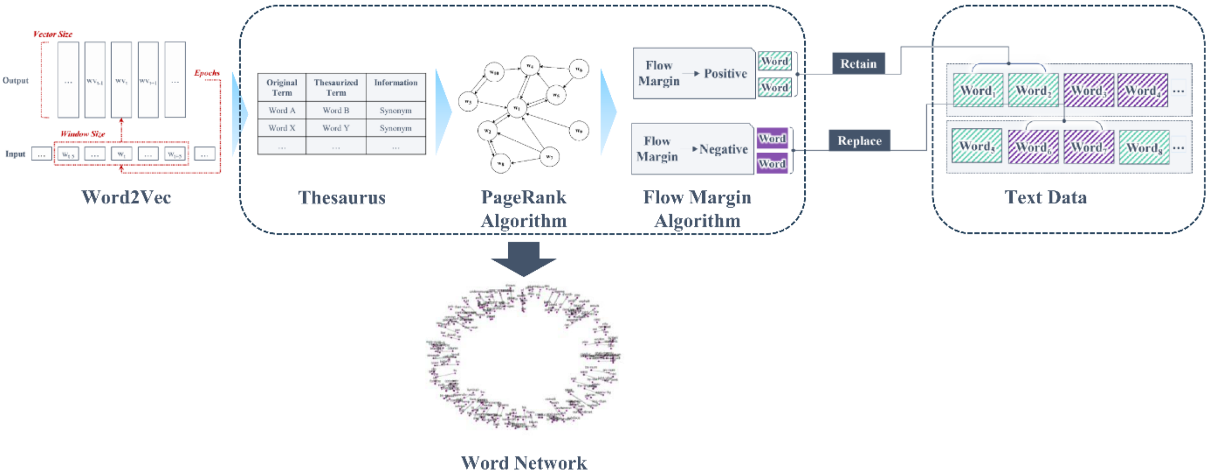


Figure 1. Research Framework

2. LITERATURE REVIEW

2.1. Thesaurus

A thesaurus delineates the relationships among terms, encompassing synonyms, hypernyms, and hyponyms, rather than providing definitions for each term [1]. Various information retrieval methodologies have harnessed thesauri to broaden queries and yield comprehensive search outcomes [2]. By substituting every term in a text with its pivot form, as listed in the thesaurus, the text analysis would be expected to reflect the analysis intention of the user. Thesauri are typically constructed by industry experts possessing deep knowledge of the terminology specific to a particular domain [3], or through leveraging existing synonym dictionaries [4].

2.2. Text Embedding: Word2Vec

Text data, inherently composed in natural language, pose analytical challenges for computers due to their inability to process non-numeric information directly. Consequently, the transformation of text into a format amenable to computational analysis is imperative. This transformation, known as the embedding process, entails the conversion of data from its natural linguistic representation into a numeric vector space, thereby facilitating subsequent computational tasks [5-6]. Post-embedding, each dataset acquires a distinct position within this vector space, represented as a numeric vector, rendering it computable. The embedding process, in essence, serves as a mechanism for feature extraction pivotal in the development of machine learning models.

Word2Vec, introduced by Mikolov et al. (2013), stands as a prominent technique for its effective capture of distributed word representations within sentences, aligning semantically similar words in proximal vector spaces. Word2Vec is bifurcated into two distinct architectures: Continuous Bag of Words (CBOW) and Skip-Gram [6-7]. The CBOW model predicts a target word based on context words within a predefined window size, whereas the Skip-Gram model, conversely, predicts context words from a target word. Notably, CBOW is recognized for its computational efficiency, while Skip-Gram is lauded for its superior inferential capacity regarding text.

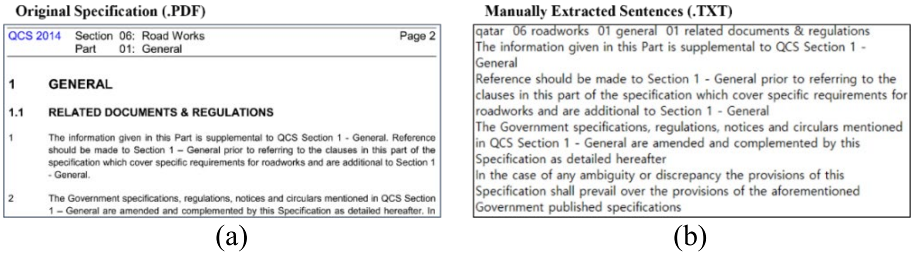
2.3. Word Weighting: PageRank

PageRank is an algorithm employed by Google's search engine, designed to allocate a weight to each document, gauging its relative significance in comparison to other interconnected documents. This weighting algorithm operates on the premise that a document of higher importance will receive a greater number of inbound links from other documents [8]. This study incorporates the PageRank algorithm to identify critical terms (i.e., pivot terms) among terms used in similar contexts. A pivot term is defined as one that serves as an effective substitute for numerous other terms (i.e., receiving a massive inflow of links) while offering few alternatives for replacement (i.e., having a minimal outflow). Within the context of the PageRank algorithm, each term is analogized to a document, with the similarities between terms represented as hyperlinks.

3. DATA

3.1. Data Collection

For this analysis, a comprehensive dataset comprising 56 construction specifications was amassed, among which two were practical specifications derived from construction projects undertaken by Korean contractors in Qatar, specifically for road construction projects in the years 2010 and 2014. The rest, constituting 54 specifications, were identified as national (or regional) standards, universally applicable to any road construction endeavor within the respective country (or region). Given the meticulous organization of standard specifications in most developed nations, the author prioritized the acquisition of recent specifications from four countries: the United States of America (USA), the United Kingdom, Canada, and Australia, sourced directly from their national or governmental websites.



Original Specification (.PDF)

QCS 2014 Section 06: Road Works Page 2
Part 01: General

1 GENERAL

1.1 RELATED DOCUMENTS & REGULATIONS

1 The information given in this Part is supplemental to QCS Section 1 - General. Reference should be made to Section 1 – General prior to referring to the clauses in this part of the specification which cover specific requirements for roadworks and are additional to Section 1 - General.

2 The Government specifications, regulations, notices and circulars mentioned in QCS Section 1 – General are amended and complemented by this Specification as detailed hereafter. In

(a)

Manually Extracted Sentences (.TXT)

qatar 06 roadworks 01 general 01 related documents & regulations
The information given in this Part is supplemental to QCS Section 1 -
General
Reference should be made to Section 1 - General prior to referring to the
clauses in this part of the specification which cover specific requirements for
roadworks and are additional to Section 1 - General
The Government specifications, regulations, notices and circulars mentioned
in QCS Section 1 - General are amended and complemented by this
Specification as detailed hereafter
In the case of any ambiguity or discrepancy the provisions of this
Specification shall prevail over the provisions of the aforementioned
Government published specifications

(b)

Figure 2. Example of data format conversion: (a)original PDF, (b)manually extracted TXT

Acknowledging the direct impact of data quality on the analytical outcomes, a manual conversion methodology was adopted, specifically the drag-copy-paste technique, meticulously extracting each sentence from the PDF documents individually (Figure 2). Consequently, this meticulous approach yielded a dataset ready for analysis, encompassing 2,527 clauses (i.e., 19,346 sentences) across six regions, including two construction specifications (i.e., QCS 2010 and QCS 2014) and four national standard specifications,ensuring comparability among them.

3.2. Text Preprocessing

This preprocessing phase encompasses several critical processes: tokenization, stopword removal, and stemming or lemmatization, each tailored to the specific objectives of the analysis [2, 5]. Tokenization dissects sentences into individual words, treating each as the fundamental unit of analysis. This research adopts an n-grams approach, recognizing clusters of adjacent tokens that exhibit frequent co-occurrence within the corpus, such as "should be," "job mix formula," and "asphalt plant." This method of multigram tokenization significantly reduces the data's feature dimensionality, thereby refining the text analysis outcomes [9-10]. N-grams manifesting at least ten occurrences across the documents, with n ranging from 2 to 5, were cataloged. The n-gram "shall be" emerged as the most prevalent, recorded 10,285 times, highlighting the extensive distribution of such n-grams, denoted by a pronounced long tail. A threshold of 100 occurrences was set for n-grams to delineate the top 481, addressing the potential for rare n-grams to obscure the clarity of individual word meanings. Stopword removal involves eliminating words unnecessary for text analysis, as they are common across documents and contribute minimally as document-specific features. Stemming reduces words to their stems, a base form unaffected by morphological alterations. For instance, "pave," "pavement," "paving," and "paved" share the stem "pav." This process aids in unifying word variants under a common stem, simplifying text analysis. Conversely, lemmatization transforms words back to their lemmas or root forms, such as converting all forms of "pave" to "pave" itself.

The research adhered to four fundamental data cleaning steps: (1) converting all text to lowercase to standardize case recognition, (2) excising noise characters generated during data conversion and encoding, (3) replacing URLs, reference names, and numbers with "LINK," "REF," and "NUM" respectively to typify rather than specify information, and (4) harmonizing unit notations across specifications, treating "percent" and "%" equivalently, for instance. These preparatory measures serve to streamline subsequent text analysis, ensuring uniformity and clarity in the processed data.

4. METHODOLOGY

This research employed the Word2Vec model and Cosine similarity on preprocessed words to identify the most similarly distributed words for each term. The Word2Vec model embedded every word to the numeric vector, then the Cosine similarities between the word vectors were calculated. The results would be a dictionary of similarly used words, of which example is provided in Table 1.

Table 1. Dictionary of similarly used words

Word	1st	2nd	3rd
w_1	w_2	w_4	w_5
w_2	w_1	w_8	
w_3	w_1	w_{10}	
w_4	w_1		
w_5	w_1	w_4	
w_6	w_4	w_5	
w_7	w_1	w_2	w_8
w_8	w_2		
w_9	w_1		
w_{10}	w_3	w_4	

The items of the dictionary can operate as nodes and edges of the word link graph. Each word in the dictionary would be a node in the word link graph. The relationships between the key-word (i.e., hash) and the value-words (i.e., elements of the data block) would be the edges that link the nodes. The data provided in Table 1 would be represented to a linked graph (Figure 3). As the w1 and w4 in Figure 7 shows, several words might be converted to each other recursively. To address this problem, the author proposed a simple but powerful algorithm based on the concept of PageRank.

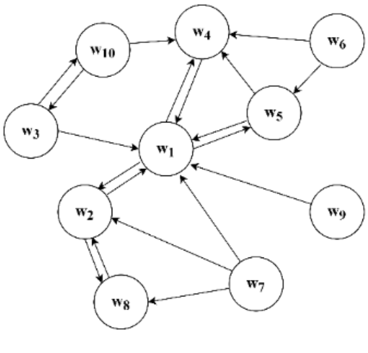


Figure 3. Example of word link graph

4.1. Word Embedding

The author employed the Word2Vec model, employing the skip-gram architecture. This model underwent training on a corpus comprising 346,950 words, equating to 8,692 distinct terms extracted from a total of 19,346 sentences, all manually curated from the collected text data. The configuration of the Word2Vec model's hyperparameters was meticulously determined based on a series of empirical analyses conducted by the researcher (Table 2).

Table 2. Hyperparameters of Word2Vec model

Hyperparameter	Value	Description
Vector Size	200	The dimension of word vector
Window Size	10	The number of adjacent words used to learn the word distribution
Minimum Count	50	The minimum frequency of each word to learn the distribution
Epochs	100	The number of iterations to learn the training data

4.2. Pivot Term Determination Based On Semantic Similarity

Based on the Cosine similarity between word vectors generated by the Word2Vec model, a dictionary of similar words was constructed. This entailed calculating the Cosine similarity for each word vector, selecting the ten most similar words for every term, and excluding those with a similarity score below 0.5. Finally, the similar word dictionary included the pairs of similar words, which facilitates the text analysis method to replace a word with its most similar word.

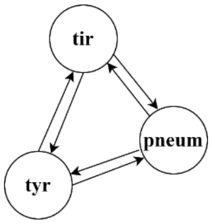


Figure 4. Sample of recursive word replacement

However, an issue emerged where certain words were recursively transformed into one another. For instance, "tyr," "tir," and "pneum" (representing the lemmas for "tyre," "tire," and "pneumatic," respectively) were found to create a recursive loop, as depicted in Figure 4. Within this network, each node represents a word, each edge denotes the relationship between two words, and the proximity between nodes inversely correlates with their Cosine similarity.

In order to address the recursive replacement problem, the authors proposed a link analysis approach of which concept is mainly based on the PageRank algorithm. Here, each word is treated akin to a document within PageRank, with the connections between words analogous to hyperlinks. According to this methodology, a term's "inflow" is defined by the count of other terms identifying it as similar. For instance, as illustrated in Figure 7, "w3" and "w4" contribute to the inflow towards "w1." Conversely, a term's "outflow" represents the number of terms it is similar to, with a similarity exceeding 0.5. For example, "w1" directs outflows towards "w2," "w4," and "w5." Terms characterized by significant inflow yet minimal outflow are deemed pivotal (i.e., pivot terms). Hence, pivot terms are ascertainable through the analysis of link numbers and flow margins, the latter of which is quantified as per Equations 1 to 4.

$$g(w', W_w) = \frac{n(W_w) - index(w'|sorted(W_w))}{n(W_w)} \quad (1)$$

$$inflow_w = \sum_{w' \in W_{w,IN}} g\left(w', W_{w,IN}\right) * c(w, w') \quad (2)$$

$$outflow_w = \sum_{w' \in W_{w,OUT}} g\left(w', W_{w,OUT}\right) * c(w, w') \quad (3)$$

$$f_w = inflow_w - outflow_w \quad (4)$$

w and w' represent individual words, while W denotes a collective set of words. Specifically, $W_{w,IN}$ is defined as the set containing word w as one of its most similar counterparts, whereas $W_{w,OUT}$ signifies the set comprising words most similar to w. The function g(w',W_w) serves as a weighting mechanism for w' within the ordered array of W_w elements, arranged according to their Cosine similarity with w. The term f_w is used to describe the flow margin of a word. Furthermore, the function c(w,w') calculates the Cosine similarity between the vectors of the input words. The algorithm of determining the pivot term is provided as below:

(1) A word is retained if its flow margin is positive.
(2) A word is substituted with alternative words if its flow margin is negative:
(2-1) Substitute the original word with its most similar word if the latter's flow margin is positive.
(2-2) If not, apply (2-1) using the subsequent most similar word.
(3) In the absence of any word with a positive flow margin, replace the original word with its closest counterpart, then proceed as per (1).

5. RESULTS AND DISCUSSION

5.1. Results of Word Embedding

The Word2Vec model generated unique vectors for 1,409 terms, indicating that the specifications commonly incorporate similar terminologies across various sentences. As an unsupervised embedding model, Word2Vec's outcomes were assessed qualitatively by the researcher.

Table 3. Samples of Word2Vec embedding result

Word	Most Similar Words
AASHTO	in-accordance-with, ASTM, standard
Contractor	by-the-contractor, proposed, his
Bituminous	asphalt, mixture, surfacing
Failure	event, load, defect
Weather	event, condition, bed

This involved the random selection of words and the examination of their most closely related counterparts. Notably, "American Association of State Highway and Transportation Officials (AASHTO)" and "American Society for Testing Materials (ASTM)," being the most frequently cited references, led to an anticipation that "ASTM" would emerge as the most similar term to "AASHTO." Contrary to expectations, the tri-gram "in accordance with" was identified as the closest match for "AASHTO," with "ASTM" ranking as the second closest. This occurrence is attributed to the frequent usage of the phrase "... in accordance with AASHTO ..." within the clauses. In a similar vein, "Contractor" was most closely associated with the tri-gram "by the contractor." The terms "Bituminous" and "Failure" yielded logical associations, with "asphalt" and "event" being their respective closest terms. Meanwhile, "Weather" was most closely linked to "event," and "condition," a seemingly more intuitive match, ranking second in similarity (Table 3). In conclusion, the Word2Vec model demonstrated commendable efficacy in capturing the distributed representation of specification texts, despite a few miss-graded similarities.

5.2 Semantic Word Similarity

Cosine similarities between word vectors were computed, and for each word, the five most similar words with similarities exceeding 0.5 were cataloged to compile a similar word dictionary. Subsequent refinement of the dictionary involved the removal of several unsuitable entries.

Table 4. Sample of similar word dictionary

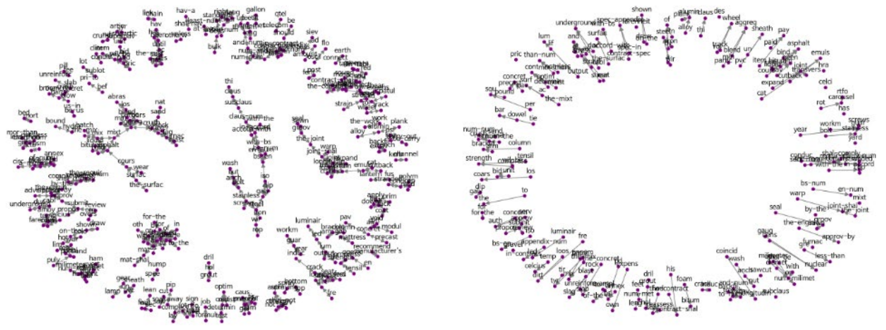
Term	Most Similar Terms				
	1st	2nd	3rd	4th	5th
asphalt	bind	mix	mixt	bitumen	liquid
aggreg	coars	mixt	crush	min	blend
mix	the-mix	mixt	asphalt	batch	-
pav	concret	-	-	-	-

Initially, 965 entries lacking similar words with a Cosine similarity above 0.5 were identified as empty data and subsequently eliminated. Furthermore, entries based on single alphabets, presumably originating from chapter titles such as "B. Asphalt Plants," were excised, totaling seven rules. Notably, the key-word “a,” associated with a data block containing prepositions like “an,” “the,” “of,” was retained as an exception. Consequently, the refined similar word dictionary comprised 374 entries (Table 4).

The structure of the similar word dictionary is depicted as a word network, illustrating directional links from the original term to its similar counterparts. Given the recurrent issue of recursive replacement observed in numerous entries, the identification of pivot terms emerges as a critical necessity (Figure 5(a)).

5.3 Semantic Construction Thesaurus

The semantic construction thesaurus was devised through the analysis of links between terms and their similar counterparts, employing a fundamental principle derived from the PageRank algorithm. This thesaurus outlines replacement guidelines (i.e., from each term to its pivot) depicted as a word network (Figure 5(b)), whereby each word at the starting points is substituted with the endpoint word across the text, with all words presented in their lemmatized forms.



(a) (b)

Figure 5. (a)Word network of similar word dictionary, (b)Semantic construction thesaurus

Table 5. Word replacement rules of semantic construction thesaurus

Index	Word		Pivot Term	
	Lemma	Original Token	Lemma	Original Token
1	temp	temperature	celcius	celcius
2	tir	tire	tyr	tyre
3	propos	propose	submit	submit
4	approv	approval	submit	submit
5	iron	iron	steel	steel
6	bitumin	bituminous	asphalt	asphalt
7	liquid	liquid	asphalt	asphalt
8	item	item	pay	payment
9	accord-with	accordance-with	with-bs	with-BS
10	shal-comply	shall comply	require-of	requirement-of

As the word replacement rules of the developed semantic construction thesaurus have no correct answers, the author evaluated the results by investigating several randomly selected rules. Table 5 provides the samples of the semantic construction thesaurus. The analysis revealed that the records indexed as 1, 2, 3, 5, and 6 were found to be logical. While some conversions might seem more intuitive if reversed (e.g., from the lemma of “temp” to “celcius”), for computational purposes, the orientation poses no issue, as the primary concern is the computer's recognition of the high similarity between the two terms. However, certain entries aim to substitute a word with its immediate successor (notably records 4, 7, 8, 9, and 10). For instance, the lemmas “accord-with” and “with-bs” originate from phrases such as “… in accordance with the BS EN ….” These inaccuracies stem from the Word2Vec model’s tendency to perceive closely positioned words as similar, a trait generally advantageous due to the high co-occurrence rate of proximate words. To refine the conversion records, several entries were manually adjusted in close collaboration with construction industry professionals to incorporate practical insights. Consequently, the construction thesaurus was finalized to include 208 conversion records.

6. CONCLUSIONS

This research aims to present a specific method for applying a thesaurus in the text mining process of construction documents using language models. The thesaurus serves as a lexicon delineating the interrelationships among terms, substituting each with a designated pivot form. Initially, the author compiled 56 construction specifications, manually transcribing 2,527 clauses (or 19,346 sentences) from the collected corpus into a TXT format suitable for text analysis. Subsequent phases involved comprehensive data cleaning and text preprocessing. Following this, the Word2Vec embedding model, utilizing the CBOW architecture, was employed to learn the distributed representations of 346,950 words (equating to 8,692 terms) extracted from the entirety of the 19,346 sentences, successfully generating unique vectors for 1,409 terms that exceeded the minimum frequency threshold. In response to the challenges posed by the similar word dictionary, which highlighted the issue of recursive word conversion, the authors devised a link analysis strategy, fundamentally inspired by the PageRank algorithm. Terms experiencing substantial inflow yet minimal outflow were identified as pivotal. Finally, the author developed the thesaurus with 208 replacement rules. Subsequently, the designed algorithm will be utilized to apply the thesaurus to language models, with the intention of investigating whether the model's performance improves as a result.

ACKNOWLEGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00241758). This research was supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 1 (RG136/22). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

[1] Aitchison. J, Gilchrist. A, Bawden. D, "Thesaurus Construction and Use: a Practical Manual", Routledge, 2003.
[2] Zou. Y, Kiviniemi. A, Jones. S. W, "Retrieving similar cases for construction project risk management using Natural Language Processing techniques.", Automation in Construction, 80, pp. 66–76, 2017.
[3] Kim. T, and Chi. S, "Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry.", Journal of Construction Engineering and Management, 145(3), 0401900, 2019.
[4] Zhang. J, and El-Gohary. N. M, "Automated Information Transformation for Automated Regulatory Compliance Checking in Construction.", Journal of Computing in Civil Engineering, 29(4), pp. 1–16, 2015.
[5] Manning. C. D, Raghaven. P, Schutze. H, "Introduction to Information Retrieval", Cambridge University Press, 2008.
[6] Mikolov. T, Chen. K, Corrado. G, Dean. J, "Efficient Estimation of Word Representations in Vector Space.", 1, pp. 1–12, 2013.
[7] Le. Q, Mikolov. T, "Distributed Representations of Sentences and Documents.", Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196, 2014.
[8] Kleinberg. J. M, "Authoritative sources in a hyperlinked environment.", Journal of the ACM, 46(5), pp. 604–632, 1999.
[9] Joulin. A, Grave. E, Bojanowski. P, Mikolov. T, "Bag of Tricks for Efficient Text Classification.", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 1–5, 2016.
[10] Wang. S, Manning. C. D, "Baselines and bigrams: Simple, good sentiment and topic classification.", 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2(July), pp. 90–94, 2012.