IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 9, 2020, accepted September 18, 2020, date of publication September 24, 2020, date of current version October 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026585

Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight

MIN-JEONG KIM[1], JI-SOO KANG[2], AND KYUNGYONG CHUNG[1]
[1]Division of Computer Science and Engineering, Kyonggi University, Suwon 16227, South Korea
[2]Department of Computer Science, Kyonggi University, Suwon 16227, South Korea

Corresponding author: Kyungyong Chung (dragonhci@gmail.com)

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 20CTAP-C157011-01).

ABSTRACT With the increase in traffic accident rates, traffic risk detection is becoming increasingly important. Moreover, it is necessary to provide appropriate traffic information considering user locations and routes and design an analysis method accordingly. This paper proposes a word-embedding-based traffic document classification model for detecting emerging risks using a quantity termed sentiment similarity weight (SSW). The proposed method detects emerging risks by considering and classifying the importance and polarity of keywords in traffic document. Conventional sentiment analysis methods fail to utilize semantically significant keywords unless they are included in a sentiment dictionary. In this study, through word imputation using an established similarity dictionary and by widening the limited utilization range, the proposed method overcomes the disadvantage of sentiment dictionaries. The proposed method is evaluated through three tests. In the first, the similarity between keywords is measured, and thus model accuracy is evaluated. In the second test, three classifiers for emerging risk classification are compared. In the last test, emerging risk detection is assessed according to whether the proposed SSW is applied, and its effectiveness is therefore verified. The evaluation results demonstrate that the proposed traffic-related document classification model using the SSW has an f-measure of 0.907, indicating satisfactory performance. Therefore, the proposed SSW can be effectively used as a parameter in traffic-related document classification and enables the detection of emerging risks.

INDEX TERMS Classification, risk detection, sentiment analysis, text mining, traffic, word embedding.

I. INTRODUCTION

The development of transportation means positively influences everyday life in several respects, such as shortening travel time and overcoming the limitations of distance travel. However, as the number of people using transportation increases, traffic volume also increases, and thus traffic congestion and accidents occur more frequently. Accordingly, the fatality rate of traffic accidents rises, and therefore the social cost for handling such accidents is increasing [1]. Traffic accidents occur unexpectedly and are difficult to analyze accurately because they are affected by environmental factors. Therefore, it is necessary to conduct long-term risk management through traffic data analysis. In addition, with the development of information and communication technology, massive amounts of unstructured data are being generated in real time through mass media and social networking services (SNS). In this circumstance, unstructured data analysis based on artificial intelligence for supporting intelligent transportation systems (ITSs) has proved quite valuable [2]. Unstructured data may be found in various forms, such as text, images, and multimedia. According to the characteristics of a dataset, it is necessary to apply different mining techniques. Among them, text and opinion mining for the analysis of unstructured text data have attracted considerable attention [3]. Opinion mining is a technique whereby useful information is extracted by analyzing people opinions. It uses sentiment information to convert the sentiment of

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

a text into quantifiable and objective information that can be analyzed. In addition, opinions are classified as positive, negative, or neutral, and this can be applied to decision making. Text mining is a technique for extracting new and meaningful information from preprocessed text data by using association rules, cluster analysis, and classification [4], [5]. Currently, text-mining-based analysis is under investigation for extracting traffic information from real-time stream text data [6]. Such data consist of a variety of text information, including words and sentences, for traffic risk assessment [7]. However, owing to the data explosion, it is difficult to extract traffic data only from massive text data. Ali *et al.* [8] conducted an ontology-based transportation sentiment analysis using unstructured text data from social network platforms to extract meaningful information. The disadvantage, however, is that sophisticated data preprocessing is essential because of the characteristics of social network content. This implies that it is difficult to assess emerging risks for ITS from simple and unprofessional traffic-related information.

This paper proposes a word-embedding-based traffic-related document classification model for detecting emerging risks using a quantity termed sentiment similarity weight (SSW). The proposed model classifies word-embedding-based traffic-related documents from news data and detects emerging risks using the SSW from the classified documents. It collects unstructured traffic data through crawling. Traffic-related main keywords are extracted from the collected data, and the importance of keywords in the document is determined by term frequency–inverse document frequency (TF–IDF) weight. By performing sentiment analysis on keywords, the polarity value of a word is determined. In this study, the SSW is proposed to resolve the issue of limited word range in sentiment dictionaries. Specifically, words are weighted considering the similarity and polarity of the main keywords. Thereby, traffic-related documents are classified, and emerging risks are detected. Accordingly, traffic-related document classification using the proposed SSW allows the detection of emerging risks considering user routes and thus provides significant information on traffic risks. This enables safe driving and walking for drivers and pedestrians, respectively. The main contributions of this study are as follows.

1. We propose a new framework that detects and classifies keywords related to emerging traffic risks by considering the polarity and importance of words using the proposed SSW.
2. We propose a method to detect emerging traffic risks by extracting only traffic-related documents from unstructured text data of various categories.
3. We overcome the limitation of sentiment dictionaries by measuring the similarity between keywords and using the proposed word imputation method.
4. We propose a graphical user interface that can detect emerging risks using the traffic-related document classification model.

The remainder of the paper is organized as follows: Chapter 2 describes sentiment word classification based on sentiment analysis, and Word2vec-based word embedding. Chapter 3 describes the proposed word-embedding-based traffic-related document classification model for detecting emerging risks using the SSW. Chapter 4 describes the experiments the results of the performance evaluation of the proposed model. Chapter 5 concludes the paper.

II. RELATED STUDIES

A. SENTIMENT WORD CLASSIFICATION BASED ON SENTIMENT ANALYSIS

Sentiment analysis is a technique for measuring the polarity value (the level of positiveness and negativeness) in a piece of text and determining the related sentimental state [9]. To this end, a sentiment dictionary with the polarity values of sentiment vocabularies is established. This is expressed by quantifying word sentiment after preprocessing text data. Sentiment analysis can be performed through sentiment-dictionary-based methods and machine-learning methods. The former analyze word sentiment by numerically representing word polarity. However, if a word is not present in a sentiment dictionary, it is difficult to analyze its sentiment. Accordingly, such dictionaries should be carefully established. To this end, supervised learning based on machine and deep learning is often applied. Supervised learning is the process of learning through user labels [10]. By labeling the level of positiveness or negativeness of a document, a classifier learns to make inferences. Sentiment-dictionary-based modules include TextBlob [11], valence-aware dictionary and sentiment reasoner [12], and SentiWordNet [13]. Madhu [14] proposed sentiment-analysis-based document clustering, which determines the polarity and subjectivity values of Twitter documents using the TextBlob and AFFIN dictionaries. Documents are clustered by sentiment analysis, and thus their relationship is discovered. Denecke [15] proposed a method for automatically determining the polarity of multilingual documents, whereby such a document is translated into English through standard translation software, and then the Sentiwordnet dictionary of emotions is used to determine its polarity. That is, a single sentiment dictionary is sufficient in a multilingual framework.

General sentiment dictionary construction consists of a text data collection step and a morphological and sentiment analysis step. The former is the collection and pretreatment of text data. In the pre-processing process, data are collected and documented, and disused words are removed. Moreover, words are tokenized. In the latter, morphological analysis and sentiment analysis are conducted on the collected text data. The documented data are analyzed morphologically and tagged, and then their sentiment is analyzed. In addition, words are analyzed in terms of sentiment level, are classified into positive, neutral, or negative types, and are quantified. The process of establishing a sentiment dictionary is shown in Fig. 1. Conventional sentiment analysis methods use the polarity values of vocabularies in a sentiment dictionary.

The degree of positiveness and negativeness can be determined only for words that appear in a sentiment dictionary.

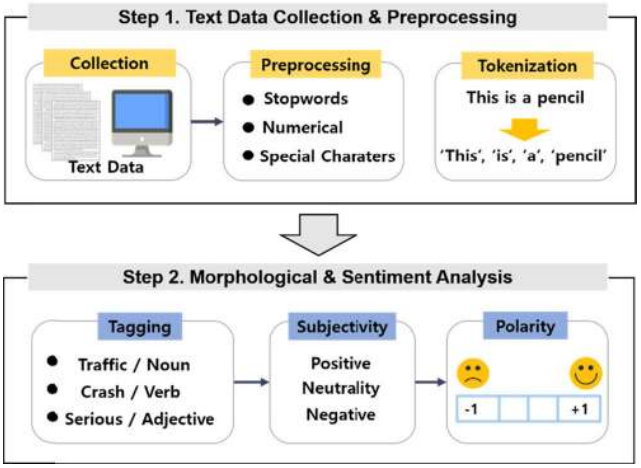

FIGURE 1. Establishing a sentiment dictionary.

Therefore, to expand a conventional sentiment dictionary, it is necessary to design a method of calculating the polarity value of missing words.

B. WORD2VEC-BASED WORD EMBEDDING

Word embedding is a technique for analyzing the context of words in a sentence and converting them into a vector value. Word embedding methods include GloVe [16], Fasttext [17], and Word2vec [18]. Word2vec fails to consider the co-occurrence frequency of entire sentences because learning is performed only in a user-specified window. By contrast, GloVe (global vectors) determines the structure of semantic words by using the co-occurrence probability of all words. However, as the number of words increases, the size (hence, the computational complexity) of the co-occurrence word matrix increases [16]. Fasttext is a word-embedding technique proposed by Facebook. It is assumed that there are multiple words in a term, to obtain the word embedding vector, Fasttext splits a word into n-grams, and the embedding vector is then the sum of these n-grams. For example, 'kill' $\rightarrow$ (ki, il, ll). This generates a vector of words that are not found in the dictionary; moreover, its training process is fast [17]. Word2vec was proposed by Google after the neural network language model was improved [18]. It infers the meaning of words based on the distribution hypothesis that words in similar locations have a similar meaning. In addition, a vector is assigned to a word by representing the word in high-dimensional coordinates [19]. As operations between words are possible, the similarity between words can be calculated. The continuous bag of words (CBOW) and skip-gram are representative Word2vec methods [20]. The former uses surrounding words as input to predict a target word and has the advantage of fast training [21]. Skip-gram takes a target word as input to predict the surrounding words. Its training process is slow because the loss must be calculated by the number of contexts [22]. Nevertheless, CBOW is inferior to skip-gram in terms of learning efficiency because it updates the vector value of a target word only once, whereas in skip-gram, the size of the context window is twice as large as the window size. Therefore, for the same window size, the learning amount can be several times as large in skip-gram. Accordingly, skip-gram with high learning efficiency is generally applied [23]. Seyed Mahdi Rezaeinia [24] proposed an improved word embedding method based on a pre-trained sentiment dictionary by applying sentiment analysis. It extracts vectors from a text corpus according to Word2vec/GloVe, a word position algorithm, a vocabulary-based approach, and morphological analysis. In combination with the extracted vectors, improved word vectors are constructed. Thereby, accurate classification can be achieved; however, the proposed method was evaluated only on sentiment datasets. B. Naderalvojoud [17] developed a sentiment-recognition-based word-embedding deep learning model. It considers both the meaning and polarity of words and overcomes the weakness that words with different sentiments, but similar contexts may have similar vector values. However, this method does not consider the importance of words in a document. Conventional sentiment analysis techniques may be unable to use meaningful words because of the limited range of sentiment dictionaries. Therefore, it is necessary to develop a method whereby used in sentiment analysis, although they may not appear in a sentiment dictionary.

III. WORD-EMBEDDING-BASED TRAFFIC-RELATED DOCUMENT CLASSIFICATION MODEL FOR DETECTING EMERGING RISKS USING SENTIMENT SIMILARITY WEIGHT

Even though it is possible to obtain information from text data, it is difficult to recognize and predict risks based on these data. This study proposes a technique for collecting traffic-related documents from a variety of texts and detecting emerging risks. The proposed technique substantially uses words that may not appear in a sentiment dictionary by applying the SSW. Figure 2 shows the word-embedding-based traffic-related document classification model for detecting emerging risk using SSW.

The proposed method consists of four steps. In the first step, collect unstructured data through crawling, preprocess it, and then form a matrix through TF-IDF weights. This removes the stop words of the collected documents and proceeds with the morpheme analysis process. It also extracts important keywords with high TF-IDF weight. In the second step, document labeling is performed based on the main keywords extracted from the document. Through this, it is classified into traffic-related data and non-traffic data. In the third step, Word2vec is applied to vectorize words, calculate a similarity value, and generate a dictionary. In the last step, the SSW is extracted using the TF–IDF, similarity, and polarity values of the keywords. Thereby, words not present in a sentiment dictionary are replaced by highly similar words. In addition, a classification model using the SSW is applied to a user interface system.

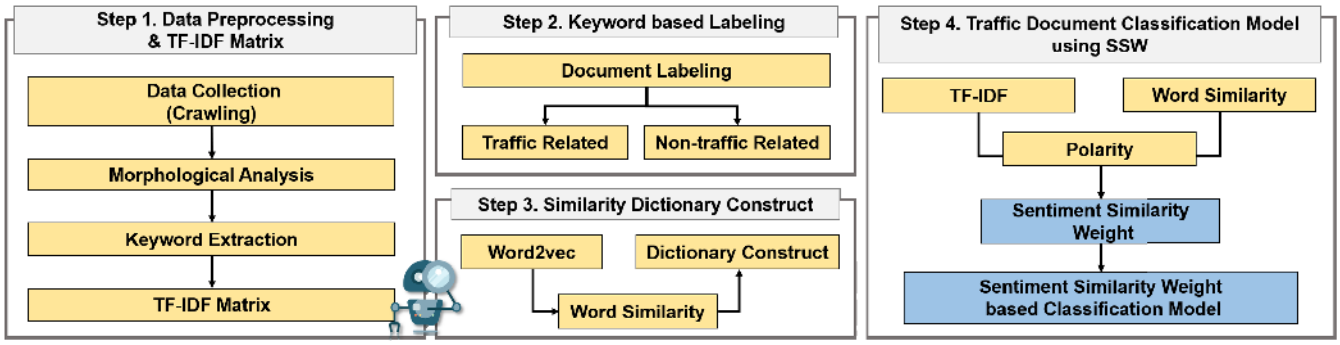

FIGURE 2. Word-embedding-based traffic-related document classification model for detecting emerging risk using SSW.

A. DATA COLLECTION AND KEYWORD EXTRACTION USING TF–IDF

To detect emerging risks in a traffic-related document, breaking news data are crawled. The data provided by the Traffic Broadcasting Network (TBN) [25] are collected and documented according to topic (e.g., politics, economy, society, information technology and science, and traffic). Subsequently, only necessary data are retained, such as date and time, category, title, main body, links. In the crawled news text data, not all categories are ''traffic,'' but there are traffic-related documents. Therefore, it is necessary to determine the category of the collected documents again. To this end, keywords are extracted. Based on the extracted keywords, binary labeling is performed to determine whether the document is traffic-related. Morphological analysis is conducted to extract representative keywords from the crawled documents. Morphological analysis is used to grasp the structure in the minimum unit of a semantic corpus [26]. Words in each document are tokenized through morphological analysis, and meaningless words are removed, that is, conjunctions, postpositions, numbers, and special characters. To extract the main keywords from the preprocessed data according to word importance, a matrix of TF–IDF weights is constructed. It multiplies the frequency of the term with the reverse document's frequency, merely reducing the weight of the many words. This allows us to extract meaningful keywords. The morphemes used for keyword extraction are nouns. However, it is difficult to conduct sentiment analysis using one part of speech, as sentiment vocabularies contain various parts of speech. Therefore, for accurate and effective sentiment analysis, verbs and adjectives that are included in sentiment vocabularies are also extracted. The extracted keywords are listed, and TF–IDF weight matrices for nouns, verbs, and adjectives are established. Table 1 shows the TF–IDF weight matrix for nouns. The top keywords are presented according to the TF–IDF weights for noun words in each document.

In Table 1, the first row indicates the pronunciation of Korean words, and their translation into English (in the parentheses). For example, the TF–IDF weight of ''sa-go'' (accident) in Doc No. 54 is 0, and its TF–IDF weight in Doc No. 128 is 0.741. That is, ''sa-go'' is meaningless in Doc No. 54, but meaningful in Doc No. 128. This is because each document has different word importance. Therefore, by constructing a matrix of TF–IDF weights, it is possible to find the main keywords in a document and calculate their importance.

TABLE 1. TF–IDF weight matrix.

NO.	sa-go (accident)	gyung-chal (police)	sool-chi-han (drunk)	wi-hum (risk)	cha-ryang (vehicle)	…
Doc_{54}	0	0.315	0	0.210	0	…
Doc_{72}	0.446	0.147	0.179	0.238	0	…
Doc_{103}	0	0.526	0	0	0.364	…
Doc_{128}	0.741	0	0	0.512	0.770	…
Doc_{139}	0.678	0	0.199	0	0.541	…
Doc_{165}	0.567	0.631	0	0.279	0.128	…
…	…	…	…	…	…	…

After the topic of a document is identified by extracting the main keywords, the label '1' is assigned to the document if it is related to traffic; otherwise, the label '0' is assigned. That is, binary labeling is applied. Document labeling is based on the traffic keywords in the land, infrastructure, and transport terminology dictionary issued by the Ministry of Land, Infrastructure, and Transport [27]. Label 1 represents a document containing traffic-related words, such as ''open highway,'' ''traffic safety law revision,'' ''traffic accident,'' and ''road congestion.'' Label 0 represents a document without traffic-related words. If Label 1 is assigned to a document (i.e., the document is related to traffic), its location is collected. A document with Label 0 is not collected but deleted. Keyword labeling is based on the actual content of a document, rather than on document category. Therefore, this labeling allows more accurate classification. Table 2 shows the extracted keywords from a document and the result of traffic labeling.

It can be seen the crawled document is preprocessed so that it has semi-structured contents: time, location, category, and label. The contents represent the words and morphemes used in the document. ''NNG'' indicates a general noun, ''NNP''

TABLE 2. Extracted keywords and traffic labeling.

NO.	Keyword	Time	Location	Category	Label
1	Gangdong-gu/NNP, eo-rin-e(children)/NNG, bo-ho(protection)/NNG, gwa-sok(overspeeding)/NNG, alarm/NNG, sin-ho(signal)/NNG, system/NNG, gang-hwa(enhance)/NNG, ppa-reu-da(fast)/VV...	2020-01-15 21:56	Gangdong-gu	Society	1
2	Incheon /NNP, ji-yeok(area)/NNP, bal-sang(occurrence)/NNG, o-to-ba-e (motorcycle) /NNG, sa-go(accident)/NNG, jeung-ga(increase)/NNG, na-ta-na-da(appear)/VV, gyo-tong(traffic)/NNG, (je-han)limit/NNG...	2020-01-31 13:26	Incheon	Traffic	1
3	a-chim(morning)/NNG, do-ro(road)/NNG, ja-dong-cha(car)/NNG, taxi/NNG, chung-dol(collision)/NNG, gyo-tong(accident)/NNG, un-haend(driving)/NNG, unjeongja(driver)/NNG, juk-da(die)/VV, byeong-won(hospital)/NNG, tap-seung-ja(passenger)/NNG, sim-gak(serious)/NNG, son-sang(damage)/NNG, e-dong(transfer)/NNG...	2020-02-19 11:27	Seobuk-gu, Cheonan-si, Chungcheongnam-do	Traffic	1
4	audio/NNP, sa-yong-ja(user)/NNP, jeung-ga(increase)/NNG, hae-woi(overseas)/NNG, audio-chaek(audiobook)/NNP, hoi-sa(company)/NNG, ttwi-eo-deul-da(dive)/VV, service/NNG, so-ri(sound)/NNG, da-yang-seong(diversity)/NNG...	2020-01-01 10:31	-	IT Science	0
5	do-ro(road)/NNP, gyo-tong(traffic)/NNG, kwon-han(authority)/NNG, dan-sok(regulation)/NNG, camera/NNG, seol-chi(installation)/NNP, jae-tam-saek(research)/NNG, go-sok-do-ro(highway)/NNG…	2020-02-07 9:40	Gyeongbu Express, Seoul toll booth	Traffic	1
6	truck/NNG, un-jeon(driving)/NNG, gyo-tong(traffic)/NNG, bul-bit(light)/NNG, yeon-swae(concatenation)/NNG, chung-dol(collision)/NNG, (Jung-bu nae-ryuk go-sok-do-ro)Central Inland Expressway /NNP, sa-go(accident)/NNG...	2019-12-24 11:34	Central Inland Expressway	Traffic	1
7	fastfood/NNG, coffe/NNP, sang-jum(store)/NNP, ja-dong-cha(car)/NNG, ta-da(ride)/VV, chaek(book)/NNG, bil-li-da(borrow)/VV, ban-nap(return)/NNG, convenient/NNG, sa-yong-ja(user)/NNP…	2020-01-01 15:14	-	Society	0
8	ji-ha-chul(subway)/NNG, bal-saeng(operation)/NNG, jung-dan(interruption)/NNG, boramae/NNP, yeok(station)/NNG, ddu-nam(departure)/NNG, won-in(cause)/NNG, mum-choo-da(stop)/VV...	2020-01-22 23:42	Boramae Station	Traffic	1
…	…	…	…	…	…

a proper noun, and ''VV'' verb. Time refers to the date and time of collection. Location represents the occurrence position and range of a traffic event found in a document. Categories refer to politics, society, economy, information technology and science, and traffic. Existing documents have the disadvantage that they are not classified as traffic documents if the category is not ''traffic'' even if their content is traffic-related. In contrast, the keyword-based method can correctly classify such documents. In addition, it is possible to collect traffic documents that are in different categories. Therefore, the proposed method can overcome the limitation of document collection in a restricted category and improve accuracy.

B. COSINE-SIMILARITY-BASED KEYWORD SIMILARITY USING WORD2VEC

A word may have multiple synonyms with different morphemes but the same meaning. If unused words are detected, it is possible to replace them by synonyms by discerning a semantic similarity. To calculate this similarity, Word2vec is applied to vectorize the extracted keywords, and a dictionary is established using the similarity results. To this end, crawled document data are used. To vectorize the keywords, Word2vec predicts a word through the context of a sentence. Therefore, stop words are not processed in establishing a similarity dictionary. In addition, to ensure diversity, not only traffic-related documents but also documents related to various fields, such as economy, politics, and society, are used. A total of 35,729 word vectors are generated. Figure 3 shows the results of word embedding based on Word2vec.

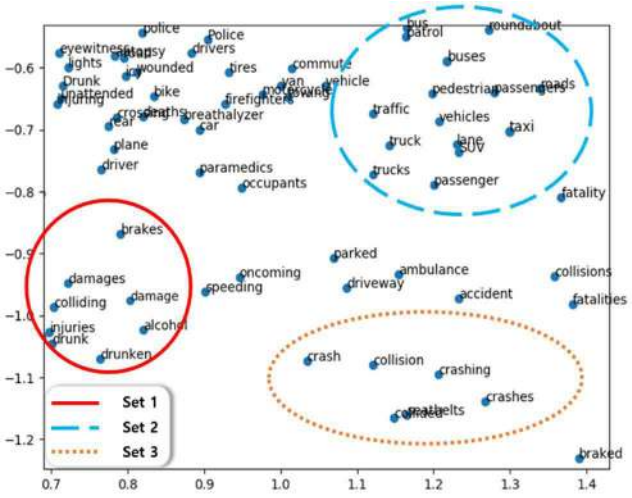

FIGURE 3. The result of word embedding by Word2vec.

It visualizes the dimensionality reduction applied to a high-dimensional vector space to obtain a two-dimensional

TABLE 3. Cosine similarity between words.

Tn	un-jeon(driving)		ja-dong-cha(car)		go-sok-do-ro(highway)		…
T1	un-haeng(race)	0.876	truck	0.913	do-ro(road)	0.884	...
T2	joo-haeng(run)	0.749	ja-dong-cha(automobile)	0.854	cha-do(roadway)	0.809	…
T3	un-jeon-ja(driver)	0.744	SUV	0.733	un-jeon-ha-da(drive)	0.751	…
T4	do-ro(road)	0.688	un-jeon-ha-da(drive)	0.724	hyu-sik(resting)	0.709	…
T5	gyo-tong(traffic)	0.632	ba-kwi(wheel)	0.768	neu-ri-da(slow)	0.692	…
…	…	…	…	…	…	…	…

space. The x- and y-axis represent the vector coordinates of a keyword.

In Fig. 3, Set1, Set2, and Set3 are the clustering sets of semantically similar keywords, representing the sets damage, crash, and vehicle, respectively. The clustered keywords are located closer to each other as the similarity increases.

To discover the similarity between words, cosine similarity is used in a vector space. It is a similarity measure between vectors and is calculated using their cosine angle. It allows the calculation of the distance between vectors in a multidimensional space [28]. Cosine similarity is a real number between -1 to 1. If the cosine similarity between two words is close to -1, then the words tend to have opposite meaning; if it is close to 1, they tend to have nearly the same meaning [29]. Equation (1) shows the formula for cosine similarity, where W and X are word embedding vectors, $S(W_k, X_k)$ denotes the cosine similarity between W_k and X_k, k means for each word and n is the number of words embedded.

$$S(W_k, X_k) = \frac{\sum_{k=1}^{n} W_k X_k}{\sqrt{\sum_{k=1}^{n} (W_k)^2} \times \sqrt{\sum_{k=1}^{n} (X_k)^2}} \quad (1)$$

Table 3 shows the cosine similarity between words by Word2vec. Tn is the similarity rank. The entries of the table are arranged in descending order of similarity.

The word that has the highest cosine similarity to ''un-jeon'' (driving) is ''un-haeng'' (race), with a similarity value of 0.876, and the word with the second highest is ''joo-haeng'' (run), the similarity value of which is 0.749. Therefore, by establishing a similarity dictionary, it is possible to obtain words that are similar to main keywords not present in a document. Moreover, word imputation is performed. That is, a word is replaced on the basis of its similarity if it is not present in a document. Nevertheless, it is possible that the worst case may occur: T1 (representing the highest cosine similarity value obtained) corresponds to a remarkably low value. Then, an original keyword is not semantically similar to T1. For example, if T1 for the word ''bi'' (rain) is the word ''ja-jeon-geo'' (bike), the similarity between the words ''bi'' and ''ja-jeon-geo'' is low (0.217). This implies that these words are semantically irrelevant. Accordingly, word imputation is applied only to words that have a similarity value higher than a threshold. For the imputation threshold setting, each of the 35,729 words is compared with T1 in terms of similarity. Whether the correlation between each word and T1 is significant is determined by decreasing the similarity by 0.1. Table 4 shows the semantic similarity probability of a word and T1 according to the imputation threshold values.

TABLE 4. Probability values according to imputation threshold.

Threshold	1.0	0.9	0.8	0.7	0.6	0.5	0.4
Value	1.0	0.875	0.812	0.794	0.628	0.211	0.193

In Table 4, the first row represents a threshold similarity value between a word and T1. The second row indicates the semantically significant probability between a word and T1 for each threshold value. It can be seen that when the threshold value is 0.5, a cutoff occurs. Therefore, two words are not considered semantically consistent if their similarity is 0.5 or less. Accordingly, the imputation threshold value is set to 0.5. Thus, through word-similarity-based imputation, it is possible to obtain words with significant correlation.

C. SENTIMENT SIMILARITY WEIGHT FOR DETECTING EMERGING RISKS

To establish the proposed SSW for detecting emerging risk, the Korean Sentiment Analysis Corpus (KOSAC) [30] sentiment dictionary is applied. It consists of the polarity values (positiveness, neutrality, and negativeness) of 16,000 n-gram morphemes. The polarity value ranges from -1 to $+1$. Figure 4 shows the extraction process of the SSW. The weight values are generated using the TF–IDF, similarity, and polarity values of the main keywords in a document. The final SSW is obtained through the sum of SSW of the words found in the document.

Regarding polarity, if a words does not appear in a sentiment dictionary, it is impossible to use it. For this reason, word imputation is applied. It considers similarity and polarity according to the existence of a word in a sentiment dictionary. Nevertheless, if the polarity of a word with the highest similarity is simply selected in the imputation process, but similarity and semantics are ignored, then the weight value may be affected. Therefore, multiply the value by Similarity to imputation only a similar degree between words. For instance, assuming that the word ''chung-dol'' (crashing) is replaced by the word ''chung-dol-ha-da'' (crash), as shown in Table 5, if similarity is not considered, the polarity value of the word ''chung-dol'' becomes -1, which is the polarity of the word ''chung-dol-ha-da''. If similarity is considered, the polarity value of the word ''chung-dol'' becomes -0.869, which is obtained by multiplying the similarity value (0.869) by the polarity value of the word ''chung-dol-ha-da'' (i.e., -1). When the similarity between two words is high, they are considered semantically similar. For this reason, the polarity value of a replacement word is significantly

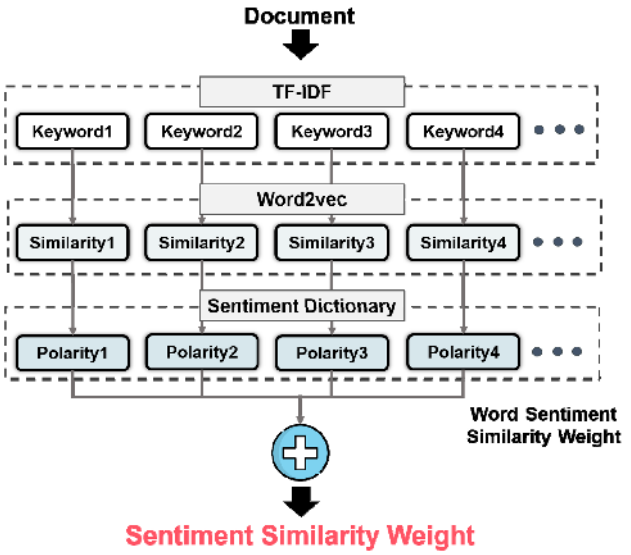

FIGURE 4. Extraction of SSW.

considered. If similarity is low, the polarity value of a replacement word is a little considered. Thus, word imputation is performed by considering the similarity between words.

The SSW for a word w ($WSSW_w$), where w may be a noun, adjective, or verb in a document, is calculated in Equation (2), where there are two cases depending on whether a word appears in a sentiment dictionary: If a keyword appears in a sentiment dictionary, only the TF–IDF and polarity values are used; otherwise, word imputation is applied using a similarity dictionary. Here, *Similarity* refers to the similarity value of the word to be replaced, *Polarity* represents the polarity value of a word.

$$WSSW_w = \begin{cases} TFIDF \times Similarity \times Polarity \\ TFIDF \times Polarity \end{cases} \tag{2}$$

For example, if the extracted word ''cha-ryang'' (vehicle) from a document is not found in a sentiment dictionary, a conventional method cannot use this word. The proposed method finds a word with high similarity to ''cha-ryang'', and the polarity value of the new word replaces that of the initial word. For instance, the polarity value of ''cha-ryang'' is replaced by the polarity value of ''ja-dong-cha'' (car), which is the most similar word. As a result, although the TF–IDF value of ''cha-ryang'' is 0.364, its polarity value is set to 0.4, which is that of the word ''ja-dong-cha''. We note that the similarity between ''cha-ryang'' and ''ja-dong-cha'' is 0.913, and thus the WSSW of ''cha-ryang'' is calculated as 0.133 in Equation (2). In Equation (3), the SSW is calculated using the WSSW values as follows:

$$SSW = \frac{1}{n} \sum_{k=1}^{n} WSSW_k \tag{3}$$

That is, the SSW is calculated by adding all the WSSW values, which are the weights of the noun, verb, and adjective keywords extracted from a document, and then by dividing the sum by the number of extracted words. In this equation, n denotes the number of words. As the number of extracted keywords differs among documents, it is necessary to divide the sum by this number to obtain a representative value. The mean of the WSSWs of all keywords is calculated, and then the SSW of Doc No. 15 is obtained. The calculation involves the top keywords, parts of speech, TF–IDF values, Word2vec-based replacement word and similarity, polarity values, and WSSWs. Table 5 shows the extraction process of the SSW. Regarding the parts of speech, NNG, NNP, and VV mean general noun, proper noun, and verb, respectively. The generated SSW ranges from −1 to +1.

TABLE 5. Extraction of SSW from Doc No. 15.

	Doc 15					
Keyword	sa-go (accident)	juk-eum (dead)	JEJU	bam (night)	bi (rain)	chung-dol (crashing)
Word Class	NNG	NNG	NNP	NNG	NNG	NNG
TF–IDF	0.883	0.705	0.687	0.591	0.584	0.582
Similarity	X	X	X	X	X	crash 0.869
Polarity	−1	−1	0	+0.5	−1	−1
WSSW	−0.883	−0.705	0	0.026	−0.584	−0.505
SSW	−0.441					

In Table 5, the keyword ''chung-dol'' (crashing) does not appear in a sentiment dictionary; thus, it is replaced by the word ''chung-dol-ha-da'' (crash), which has the highest similarity value in a similarity dictionary. In the SSW extraction process, the SSW of Doc No. 15 is calculated as −0.441.

D. TRAFFIC-RELATED DOCUMENT CLASSIFICATION MODEL USING SENTIMENT SIMILARITY WEIGHT

In this study, to detect traffic emerging risks, the proposed SSW is applied to a support vector machine (SVM) classifier [31]. Considering the polarity and importance of a word in a document, the keywords related to traffic emerging risk are detected and classified. In the classification, there are two classes: emerging risk and non-emerging risk. The class compares and labels keyword similarities between documents. To this end, traffic documents closely related to traffic safety, such as traffic accidents and traffic jams, are collected. Compare the similarity between the collected documents and the train, test data set documents. The cosine similarity determines this according to the TF-IDF weight matrix in each document. The cosine similarity comparison between documents specified to have similar keywords indicates an average of 0.71. Therefore, training and test datasets with similar cosine degrees of 0.71 or higher with collected traffic documents are labeled as an Emergency Risk, C_0. The non-emerging risk, C_1, is labeled through comparisons between documents, such as highway opening and traffic safety law revision, as shown in the above method. In the learning process of the classification model, the main keywords of a traffic-related document are extracted. Based on the TF–IDF, similarity, and polarity values of the extracted keywords, WSSWs are calculated, and are subsequently averaged to

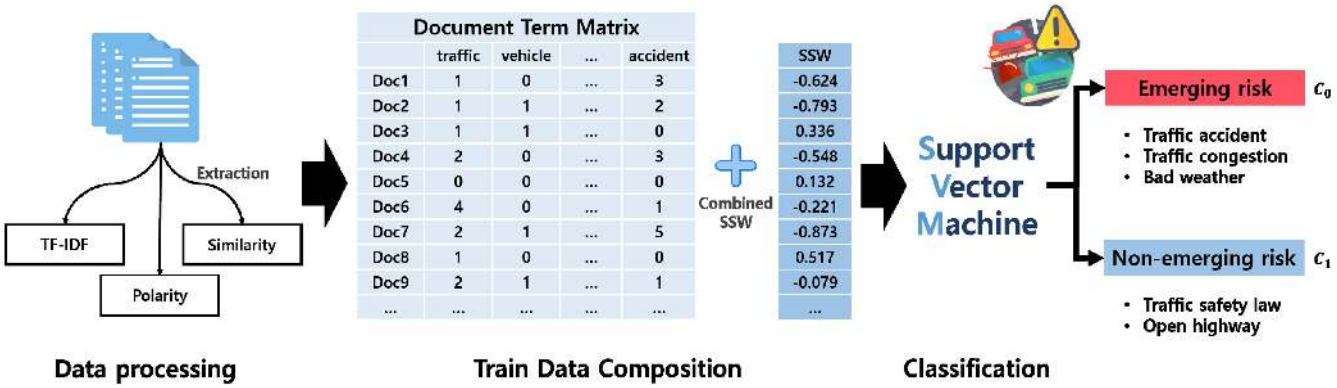

FIGURE 5. Document classification considering the SSW.

obtain the SSW of the document. Matrix is built through the keyword frequency and SSW of each document. Classification is performed by applying this to SVM. An SVM based binary classifier is used to judge if a particular document has emerging risk. Fig. 5 shows the document classification process in consideration of the SSW. Using the keyword frequency of the SSW of each document, an SSW matrix is constructed and applied to the SVM for classification. Figure 5 shows the document classification process considering the SSW.

All the WSSWs of keywords are calculated through TF-IDF, Similarity, Polarity, and with the use of the mean of the WSSWs, an SSW is calculated. In addition, the calculated SSW is combined with DTM (Document Term Matrix) to form training data. The training data considering SSW is trained in the SVM binary classifier. Considering SSW in the proposed classification model learning process, it acts as a measure to classify traffic safety-related documents and detect the risk of emergence.

IV. RESULTS AND PERFORMANCE EVALUATION

A. TRAFFIC-RELATED DOCUMENT CLASSIFICATION BASED ON EMERGING RISK DETECTION SYSTEM

An emerging risk detection system applies the proposed SSW-based classification model to a real traffic situation. It is established based on a user route. The system detects the emerging risks in this route and provides related information. We use a computer with Intel (R) Core (TM) i5-3570 at 3.40 GHz, and 16 GB RAM, running Windows 10 and Python 3.6.0. Figure 6 shows the user route-based emerging risk detection system to which the SSW-based classification model is applied.

It can be seen that the user interface of the emerging risk detection system consists of the following parts: crawling, keyword extraction, location, and emerging risk detection. It sets the url, date, and page, and then starts crawling. In addition, documents are selected, and stop words are removed. Morphological analysis is performed on documents processed with stop words. After the TF–IDF values are obtained from pre-processed documents, classification is performed using an SSW-based SVM. In the location part, the system receives the user departure and destination points, and displays on a map the number of emerging risk documents related to the route. In addition, based on the emerging risk keywords extracted from the documents, it is possible to provide the user with simple information on each route section.

Figure 6 shows the emerging risk detection mechanism applied on a route from the Gyeong-gi Provincial Government to Gwang-myeong City Hall. On the map, 265 emerging risk documents are detected in four sections. Among the documents, 132 documents are detected in the Suwon-Gwang-myeong Highway section, which is considered to carry the highest risk. The keywords detected from the extracted documents are displayed, and therefore it is possible to determine the risk factors in any road section. By providing information regarding overall and emerging risks on a route, users may prepare accordingly.

B. PERFORMANCE EVALUATION

To evaluate the performance of the proposed system, TBN news data were crawled. A total of 1,400 data documents were collected and divided into training (70%) data and test data (30%). The performance evaluation has three objectives. The first objective is to determine the best similarity evaluation method when the similarity between keywords is measured. It is evaluated by comparing the accuracy of models according to the method of measuring the similarity between words. The second objective is to select the most appropriate classification model for detecting an emerging risk by applying the proposed SSW. To this end, the proposed SSW is applied to the SVM, the KNN [32], and the naive-Bayes classification model [33], and the performance of these models in classifying emerging risk documents is compared. The third objective is to demonstrate the effectiveness of emerging risk document detection based on the proposed SSW. That is, based on an SVM binary classifier, the model with the proposed SSW is compared with a conventional model without the proposed SSW.

In this study, word similarity is measured by the cosine similarity in a vector space. Regarding the first objective,

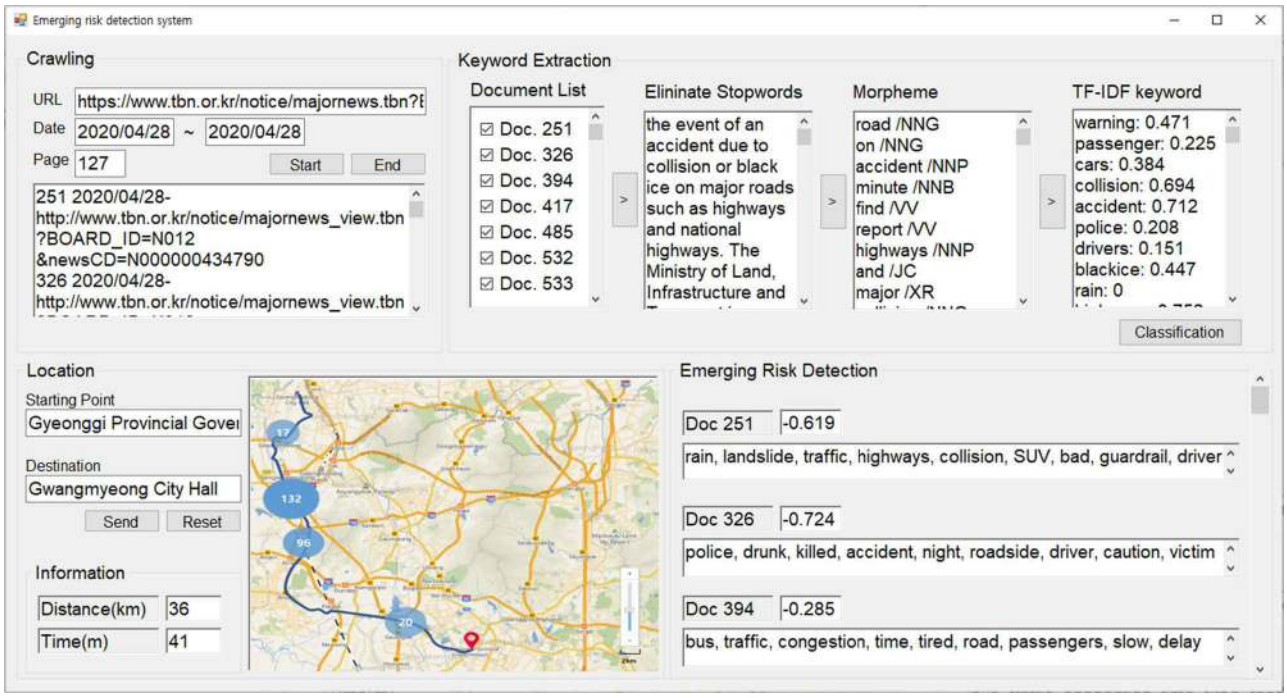

FIGURE 6. User-route-based emerging risk detection system using traffic-related document classification model.

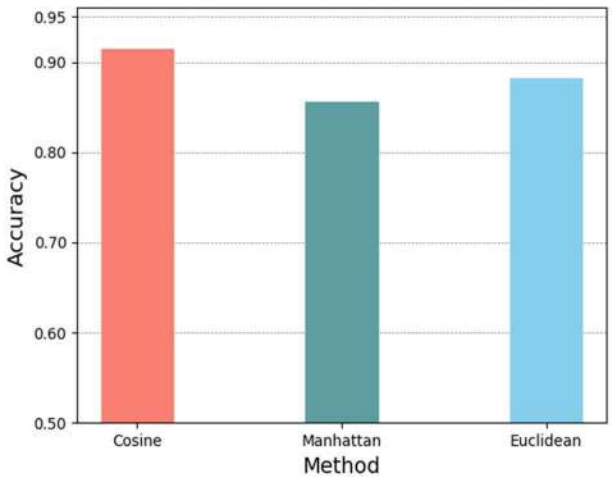

FIGURE 7. Comparison of accuracy according to similarity measure.

the cosine-similarity-based method is demonstrated to be the best similarity measurement technique. To evaluate document classification accuracy, the cosine similarity measure [34] is compared with the Manhattan-distance-based similarity measure [35] and the Euclidean-distance-based similarity measure [36]. Figure 7 shows the comparison results. In this figure, the X-axis shows the similarity measurement method, and the Y-axis represents accuracy.

It is seen that cosine similarity is more accurate than Manhattan-distance-based and Euclidean-distance-based similarity by 0.059 and 0.033, respectively. Cosine similarity is applied in a multi-dimensional space and considers vector directions. Therefore, it is possible to measure similarity by considering the overall context of a keyword in a document. By contrast, distance-based similarity measures, such the Manhattan and Euclidean methods, have considerable limitations and are not suitable for natural language processing. They infer that documents in the same context are different by simply considering the frequency of words. For this reason, these distance-based similarity measures have lower accuracy than the cosine similarity measure. Given that word embedding is performed through the contextual meanings of words, the cosine similarity technique, which considers the direction of each word vector, is more appropriate and can be used to infer the semantic similarity between keywords.

In the second performance evaluation, the proposed method proposed is applied to different classification models to select the best model. As classification models, we use SVM, KNN, and naive-Bayes. These three classifiers are compared in terms of accuracy, precision, recall, and F-measure. Precision is the probability of actual emerging risks among the predicted emerging risks [37], [38]. Recall is the probability of accurately classifying emerging risk documents. If the data are imbalanced, the scale of accuracy is unstable. For this reason, the F-measure based on precision and recall, along with accuracy, is adopted as the scale for performance evaluation [37]. Figure 8 shows the comparison results. The F-measure is the harmonic mean of precision and recall, and it is used to determine whether the classification is correct. It is given by Equation (4).

$$\mathrm{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

As shown in Fig. 8, when the proposed method is applied, the SVM classification model achieves the best performance

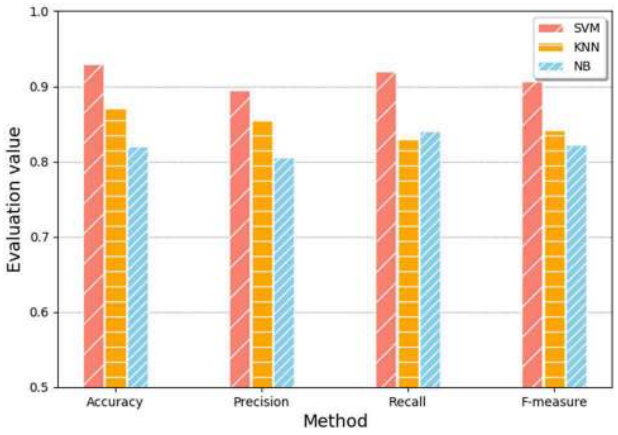

FIGURE 8. Comparison of accuracy, precision, recall, and F-measure with SSW applied to various classification models.

in terms of accuracy, precision, recall, and F-measure. This model has consistently good performance even with a small dataset and is appropriate for high-dimensional data. It is suitable because the training data has high-dimensional attributes including multiple keywords, and consists of approximately 1,400 small data points. The NB classification model does not consider the importance of keywords but views them as independent factors; thus, its performance is poor. Although KNN is rather accurate, it should calculate the similarity values of all keywords. Therefore, the method has high computational cost. Consequently, the SVM model using the proposed SSW is the best classifier for mining text, including high-dimensional keywords.

In the third performance evaluation, the SVM classifier using the proposed SSW is compared with the SVM classifier without SSW in terms of accuracy, precision, recall, and F-measure. This comparison demonstrates the effectiveness of the proposed SSW for detecting emerging risks. The results of the performance evaluation are visualized in the ROC curve graph. Table 6 shows the results of the comparison between an SVM model based on SSW and a conventional SVM model without SSW.

TABLE 6. Comparison of SSW-based SVM and conventional SVM.

	Evaluation Method	Value
SSW-based SVM	Accuracy	0.930
	Precision	0.895
	Recall	0.920
	F-measure	0.907
SVM	Accuracy	0.812
	Precision	0.805
	Recall	0.840
	F-measure	0.822

According to the overall performance evaluation, the accuracy value of the classification model based on the proposed SSW is higher than that of a conventional classification model by 0.118, and the F-measure value of the model is also higher by 0.085. By using the SSW, the proposed method effectively replaces words not present in a sentiment dictionary. The proposed technique considers the importance of words and replaces a word using its similarity even if the word does not appear in a sentiment dictionary. For this reason, it exhibits superior performance. For classification, a conventional SVM classification model without SSW excludes all the words not found in a sentiment dictionary. Such a model does not consider polarity, and therefore it exhibits poor performance. Given the results of the performance evaluation, the proposed SSW enables the effective assignment of a weight to a word (regardless of whether the word appears in a sentiment dictionary). To visualize the evaluation results, a receiver operating characteristic (ROC) curve graph is used. In the ROC curve graph, the X-axis indicates the false positive rate (FPR), and the Y-axis represents the true positive rate (TPR). The area under the curve (AUC) is the area under the ROC curve and ranges from 0 to 1. Values of AUC close to 1 indicate better performance of the model. Equations (5) and (6) define TPR and FPR, respectively, using a confusion matrix. In Equation (5), TP indicates that the classifier has determined the emerging risk document as an emerging risk, and FN indicates that the emerging risk document has been determined as a non-emerging risk. In Equation (6), FP indicates that a non-emerging risk document has been determined as an emerging risk, and TN indicates that a non-emerging risk document has been determined as a non-emerging risk.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{6}$$

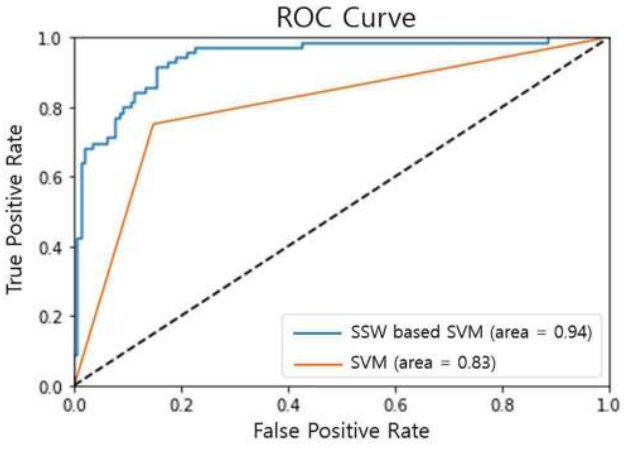

FIGURE 9. ROC curve comparison.

Figure 9 shows the results of the ROC curve comparison between a conventional SVM model without SSW and an SVM model with SSW.

It can be seen that the SVM model with SSW and the conventional SVM model without SSW have an AUC of 0.94 and 0.83, respectively. Therefore, the model using SSW has better classification performance and can effectively detect emerging traffic risks.

V. CONCLUSION

In systems using traffic data, the utilization of unstructured data is low, and data analysis and processing are more important. Thus, in this paper, we proposed a word-embedding-based traffic document classification model to detect emerging risks using the SSW. The model uses unstructured text data for analysis and processing. Specifically, it collects unstructured traffic data from breaking news. The collected data are labeled, and only traffic-related documents are used. After these documents are analyzed morphologically, the main keywords are extracted using TF–IDF weight. Based on the extracted keywords, a similarity dictionary is established by Word2vec. In addition, the similarity between words is measured. To obtain the proposed SSW based on a sentiment dictionary, the model uses the polarity, TF–IDF, and similarity values of words. If a word is not found in the sentiment dictionary, it is replaced by a word with the highest similarity using the similarity dictionary. This method overcomes the limitation of the conventional technique in which, if a particular word is not found in a sentiment dictionary, its sentiment is not extracted. In addition, the proposed model is designed to correctly apply word weights by replacing only words the similarity value of which is higher than a certain threshold. The performance of the proposed method was evaluated in three ways: Evaluate the accuracy of the model according to the similarity measurement method. Then, the proposed SSW was applied to SVM, KNN, and naive-Bayes classifiers, and their performance was compared. Finally, an SVM classifier with the proposed SSW was compared with a conventional SVM classifier without SSW. The results demonstrated that the F-measure of the traffic document classification model using the proposed SSW was 0.907, indicating good performance. Therefore, the proposed model can effectively classify traffic documents, and the SSW operates with significant parameter. In addition, by using the proposed model, an emerging risk detection system based on user route was constructed. It allows visualizing emerging risks detected from news data in a user route, thereby enabling the user to select a safe road. Using the proposed technique, it is possible to extract highly relevant traffic-related documents from unstructured data in various situations, classify the extracted documents, detect emerging risks, and notify users of potential traffic risks.

REFERENCES

[1] (Aug. 2020). *Statistics Korea*. [Online]. Available: http://kostat.go.kr/

[2] O. V. Berkout, A. J. Cathey, and K. K. Kellum, "Scaling-up assessment from a contextual behavioral science perspective: Potential uses of technology for analysis of unstructured text data," *J. Contextual Behav. Sci.*, vol. 12, pp. 216–224, Apr. 2019.

[3] Y. Jiang, D. Tao, Y. Liu, J. Sun, and H. Ling, "Cloud service recommendation based on unstructured textual information," *Future Gener. Comput. Syst.*, vol. 97, pp. 387–396, Aug. 2019.

[4] J. Xu, Y. Cai, X. Wu, X. Lei, Q. Huang, H.-F. Leung, and Q. Li, "Incorporating context-relevant concepts into convolutional neural networks for short text classification," *Neurocomputing*, vol. 386, pp. 42–53, Apr. 2020.

[5] J.-C. Kim and K. Chung, "Associative feature information extraction using text mining from health big data," *Wireless Pers. Commun.*, vol. 105, no. 2, pp. 691–707, Mar. 2019.

[6] F. Ali, D. Kwak, P. Khan, S. M. R. Islam, K. H. Kim, and K. S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 33–48, Apr. 2017.

[7] B. Alkouz and Z. Al Aghbari, "SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102139.

[8] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K. H. Kim, and K.-S. Kwak, "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowl.-Based Syst.*, vol. 174, pp. 27–42, Jun. 2019.

[9] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[10] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020.

[11] S. Kunal, A. Saha, A. Varma, and V. Tiwari, "Textual dissection of live Twitter reviews using naive bayes," *Procedia Comput. Sci.*, vol. 132, pp. 307–313, 2018.

[12] C. J. Hutto and E. E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs Social Media*, 2014, p. 82.

[13] *Sentiwordnet*. Accessed: Aug. 3, 2020. [Online]. Available: http://www.sentiwordnet.isti.cnr.it/

[14] S. Madhu, "An approach to analyze suicidal tendency in blogs and tweets using sentiment analysis," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 6, no. 4, pp. 34–36, Aug. 2018.

[15] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proc. IEEE 24th Int. Conf. Data Eng. Workshop*, Apr. 2008, pp. 507–512.

[16] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Inf. Sci.*, vol. 471, pp. 216–232, Jan. 2019.

[17] B. Naderalvojoud and E. A. Sezer, "Sentiment aware word embeddings using refinement and senti-contextualized learning approach," *Neurocomputing*, vol. 405, pp. 149–160, Sep. 2020, doi: 10.1016/j.neucom.2020.03.094.

[18] T. Ruas, C. H. P. Ferreira, W. Grosky, F. O. de França, and D. M. R. de Medeiros, "Enhanced word embeddings using multi-semantic representation through lexical chains," *Inf. Sci.*, vol. 532, pp. 16–32, Sep. 2020.

[19] N. Alami, M. Meknassi, and N. En-nahnahi, "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning," *Expert Syst. Appl.*, vol. 123, pp. 195–211, Jun. 2019.

[20] L. Niu, X. Dai, J. Zhang, and J. Chen, "Topic2 Vec: Learning distributed representations of topics," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Oct. 2015, pp. 193–196, doi: 10.1109/ialp.2015.7451564.

[21] J. Lin and W. Dongbo, "Automatic extraction of domain terms using continuous bag-of-words model," *Data Anal. Knowl. Discovery*, vol. 32, no. 2, pp. 9–15, Mar. 2016.

[22] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013, *arXiv:1309.4168*. [Online]. Available: http://arxiv.org/abs/1309.4168

[23] S. Bartunov, D. Kondrashkin, A. Osokin, and D. Vetrov, "Breaking sticks and ambiguities with adaptive skip-gram," in *Proc. Conf. Artif. Intell. Statist.*, 2016, pp. 130–138.

[24] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Syst. Appl.*, vol. 117, pp. 139–147, Mar. 2019.

[25] *Traffic Broadcasting Network*. Accessed: Aug. 3, 2020. [Online]. Available: https://www.tbn.or.kr/

[26] J. Kapusta, P. Hájek, M. Munk, and K. Benko, "Comparison of fake and real news based on morphological analysis," *Procedia Comput. Sci.*, vol. 171, pp. 2285–2293, Feb. 2019.

[27] Ministry of Land. *Infrastructure and Transsport of Korea*. Accessed: Aug. 3, 2020. [Online]. Available: http://www.molit.go.kr/portal.do/

[28] H. T. Nguyen, P. H. Duong, and E. Cambria, "Learning short-text semantic similarity with word embeddings and external knowledge sources," *Knowl.-Based Syst.*, vol. 182, Oct. 2019, Art. no. 104842, doi: 10.1016/j.knosys.2019.07.013.
[29] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2 Vec model analysis for semantic similarities in english words," *Procedia Comput. Sci.*, vol. 157, pp. 160–167, 2019.
[30] H. P. Shin, M. H. Kim, Y. M. Jo, H. Y. Jang, and A. Cattle, "Annotation Scheme for Constructing Sentiment Corpus in Korean," in *Proc. 26th Pacific Asia Conf. Lang., Inf. Comput.*, 2012, pp. 181–190.
[31] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2002.
[32] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 19–25, 2017.
[33] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naive Bayes classifier," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 99–104.
[34] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.* Berlin, Germany: Springer, Oct. 2013, pp. 611–618.
[35] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, Apr. 2013.
[36] V. M. K and K. K, "A survey on similarity measures in text mining," *Mach. Learn. Appl., Int. J.*, vol. 3, no. 1, pp. 19–28, Mar. 2016.
[37] J.-S. Kang, J.-W. Baek, and K. Chung, "PrefixSpan based pattern mining using time sliding weight from streaming data," *IEEE Access*, vol. 8, pp. 124833–124844, 2020.
[38] W. Kim and J. Chun, "An improved approach for 3D hand pose estimation based on a single depth image and Haar random forest," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 8, pp. 3136–3150, 2015, doi: 10.3837/tiis.2015.08.023.
[39] H. Kim, E. Lee, J. Chun, and K. P. Kim, "A SCORM-based e-Learning process control model and its modeling system," *KSII Trans. Internet Inf. Syst.*, vol. 5, no. 11, pp. 2121–2142, 2011, doi: 10.3837/tiis.2011.11.014.

MIN-JEONG KIM received the bachelor's degree from the Division of Computer Science and Engineering, Kyonggi University, South Korea, in 2020. She has been a Researcher with the Data Mining Laboratory, Kyonggi University. Her research interests include data mining, deep learning, and recommendation systems.

JI-SOO KANG received the B.S. degree from the Division of Computer Science and Engineering, Kyonggi University, South Korea, in 2020. She is currently pursuing the master's degree with the Department of Computer Science, Kyonggi University. She has been a Researcher with the Data Mining Laboratory, Kyonggi University. Her research interests include data mining, knowledge engineering, human-inspired artificial intelligent, data management, context deep learning, knowledge mining process, multimodal health block deep learning, and recommendation.

KYUNGYONG CHUNG received the B.S., M.S., and Ph.D. degrees from the Department of Computer Information Engineering, Inha University, South Korea, in 2000, 2002, and 2005, respectively. He has worked for the Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a Professor with the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a Professor with the Division of Computer Science and Engineering, Kyonggi University, South Korea. He was named a 2017 Highly Cited Researcher by Clarivate Analytics. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.