IEEE Open Journal of
Intelligent Transportation
Systems

Received 13 September 2023; revised 30 October 2023 and 19 November 2023; accepted 20 November 2023. Date of publication 23 November 2023; date of current version 15 December 2023.

Digital Object Identifier 10.1109/OJITS.2023.3335817

Text Classification Modeling Approach on Imbalanced-Unstructured Traffic Accident Descriptions Data

YOUNGHOON SEO[1], JIHYEOK PARK[2], GYUNGTAEK OH[1], HYUNGJOO KIM[1], JIA HU[3] (Member, IEEE), AND JAEHYUN SO[2] (Member, IEEE)

[1]Intelligent Transportation System Laboratory, Advanced Institute of Convergence Technology, Suwon 16229, Republic of Korea

[2]Department of Transportation System Engineering, Ajou University, Suwon 16499, Republic of Korea

[3]College of Transportation Engineering, Tongji University, Shanghai 200070, China

CORRESPONDING AUTHOR: J. SO (e-mail: jso@ajou.ac.kr)

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) Grant funded by the Ministry of Land, Infrastructure, and Transport under Grant RS-2021-KA162184.

ABSTRACT The unstructured-textual crash descriptions recorded by police officers is rarely utilized, despite containing detailed information on traffic situations. This lack of utilization is mainly due to the difficulty in analyzing text data, as there is currently no innovative methodology for extracting meaningful information from it. Given limitations and challenges in analyzing traffic crash descriptions, this study developed a methodology to classify significant words in unstructured data that describe traffic crash scenarios into standardized data. Ultimately, a natural language processing technique, specifically a bidirectional encoder representation from transformer (BERT), was used to extract meaningful information from crash descriptions. This BERT-based model effectively extracts information on the exact collision point and the pre-crash vehicle maneuver from crash descriptions. Its practical approach allows for the interpretation of traffic crash descriptions and outperforms other natural language processing models. Importantly, this method of extracting crash scene information from traffic crash descriptions can aid in better comprehending the unique characteristics of traffic crashes. This comprehension can ultimately aid in the development of appropriate countermeasures, leading to the prevention of future traffic crashes.

INDEX TERMS Traffic crash descriptions, natural language processing, BERT, text classification, traffic safety.

I. INTRODUCTION

TRAFFIC safety researchers and engineers frequently use traffic crash data as the core dataset describing the causes and situations of traffic crashes, which are otherwise rarely known after the event. Many traffic safety studies have analyzed traffic crash data for vehicle-to-vehicle crashes as well as vehicle-to-pedestrian crashes, motorcycle crashes, and illegal behavior [1], [2], [3], [4], [5], [6], [7], [8]. In most countries, a traffic crash report should be written after a reported crash by a responsible person whose qualifications are defined by the respective law and traffic regulations [9], [10], [11]. For example, in the Republic of Korea, traffic police officers are authorized to investigate traffic crashes according to procedures established by the national law. Traffic crash reports describe the time and place of the crash, the persons or road facilities involved, the severity and damage of the crash, violations of the law, and other external factors that may have contributed, such as weather and road conditions. While the majority of data in the crash report is obtained from closed-ended questions and a numerical coding system, certain data classes contain text-based information provided by a crash investigator. For instance, one example of such text-based data entails the following event: "Vehicle A made a left

The review of this article was arranged by Associate Editor Xin Li.

turn at an unsignalized intersection; Vehicle B approached from the opposite direction and attempted to traverse the intersection; and Vehicle A collided with the left rear bumper of Vehicle B due to failing to perceive its presence." Such elements describe crash incidents or their surroundings from the investigator's viewpoint using only factual information. Many traffic safety studies have used the structured data elements of crash reports, as they can be translated easily into statistics and are well-suited to simple data analysis [1], [2], [3], [4], [5], [6], [7], [8]; however, text-based crash descriptions are often ignored, although they reflect details that are not captured in simple numerical and code-based data. This is primarily because the nature of text renders such unstructured data difficult to manage. The limited use of crash descriptions in traffic safety research is largely attributed to the challenge in analyzing textual data due to the lack of cutting-edge techniques for extracting valuable insights from it.

However, in recent years, several studies have incorporated unstructured textual data using advanced natural language processing (NLP) techniques, aiming to identify the causes and surroundings of traffic crashes [12], [13], [14], [15]. NLP is a discipline in the field of computer science that studies the interaction between computers and human language [16], enabling computers to understand and analyze human language according to certain given rules [17], [18]. Typical examples of NLP include checking whether mail is spam, similarity-based information searches, machine translation, and chatbot generation [14], [15]. Rule-based NLP is easy to implement but requires extensive data preprocessing and model development. Recently, artificial intelligence has been applied to this issue, and artificial neural network-based models have been developed to perform high-level NLP [10], [11]. NLP techniques are commonly used in medical research and have more recently been applied in traffic studies, e.g., for developing automated vehicle scenarios using text-based traffic crash descriptions [12], [13]. However, these studies have generally used only the frequency of certain words and/or the serial connections of keywords. Therefore, they are unable to fully interpret traffic crash situations and extract implications or further information.

Hence, this study aims to develop a methodology for interpreting and extracting meaningful information from text-based traffic crash descriptions using a text-mining technique. For data interpretation and classification, this study proposes a bidirectional encoder representation from transformer (BERT) model, which is one of the most widely used and high-performing transformer models in the field of NLP [19]. This model aims to extract information from unstructured traffic crash descriptions in order to reconstruct the situation immediately prior to the crash. The innovation of this study lies in its analysis framework that extracts vital information from crash descriptions through BERT, enabling the extraction of valuable data from text-based accounts of traffic crashes. This framework for analyzing crash descriptions will serve as a valuable reference for using text-based crash descriptions, ultimately leading to a more comprehensive comprehension of events preceding and during crashes. As such, this study aims to contribute to traffic safety research in the following ways.

- Recognizing the utility of text-based traffic crash description data among the many data elements in traffic crash reports.
- Proposing a practical methodology for traffic safety engineers and crash investigators to extract new information from text-based traffic crash description data, thereby aiding the identification of crash causes and circumstances.
- Exploring the potential for understanding comprehensive crash situations based on synergy with structured traffic crash records.

II. LITERATURE REVIEW

An extensive literature review was conducted to extract key points to guide the details of this study. Although a machine learning approach has been applied for various purposes such as the traffic congestion estimation [20], [21], [22], behavior prediction [23], [24], vehicle classification [25], [26], safety estimation [27], [28], [29], and object detection [30], [31], [32], it has been relatively less used in the field of traffic safety due to the nature of text data in traffic

In the field of traffic safety, significant research effort has been devoted for using the structured data elements of traffic crash reports [1], [2], [3], [4], [5], [6], [7], [8]. Although several recent studies have attempted to utilize the unstructured text-based traffic crash description data to identify the situation and circumstances of traffic crashes, the utility of these existing methods remains limited to counting the frequency of certain keywords [12], [13], [14], [15]. As a representative example, Park et al. [14] utilized NLP technology and general automobile traffic crash data managed by the Korea National Policy Agency to develop an automated vehicle test scenario derivation methodology for safety evaluations and assurances. They generated 16 virtual urban arterial roads and 38 virtual urban intersection scenarios.

Other recent studies have used text-mining techniques in machine learning to classify crash types by interpreting the textual descriptions in traffic crash reports. Goh and Ubeynarayana [33] used crash data from the U.S. Occupational Safety and Health Administration to evaluate the usefulness of various text-mining classification techniques. They evaluated six machine learning algorithms: support vector machine (SVM), linear regression, random forest, k-nearest neighbor, decision tree, and naïve Bayes, with SVM exhibiting the highest performance in the classification of 251 test sets. Similarly, Gao and Wu [34] examined the performance of text-mining techniques in the classification of crash types based on crash reports. They employed a verb-based text-mining technique to extract and

classify syntactic and semantic word units from Missouri traffic crash records, revealing that the resulting extracted and classified information was useful in identifying causes and causal patterns of the crashes and classifying crash types. Meanwhile, Mujalli et al. [35] proposed various new methods of preprocessing to remove dataset imbalances. They argued that a three-year traffic crash dataset collected in Jordan was a disproportionate dataset, with cases of mainly minor injuries being overly represented compared to those of deaths or serious injuries. To mitigate this dataset imbalance, they used undersampling, which removes certain instances from over-represented classes, and oversampling, which generates new instances for under-represented classes. In other aspects, the NLP technique has also been used to identify the traffic situation based on the social media data in various traffic studies: identifying the traffic congestion situations [36], analyzing the mobility environment [37], incident detection [38], and designing the traffic management systems [39].

In addition, many new algorithms have been developed in the drive to advance its text classification performance. Examples include word embedding techniques and deep learning-based algorithms. Word embedding can be implemented through various algorithms, with the most common and widely used examples being term frequency–inverse document frequency (TF–IDF) [40], Word2Vec [41], FastText, and one-hot encoding [42]. These algorithms have significantly aided the achievement of high-accuracy word extraction and meaningful categorization of word units [43], [44], [45]. As the performance of graphical processing units (GPUs) has improved, deep learning models, such as convolutional neural networks (CNNs) [46] and long short-term memory (LSTM) [47], have gained prominence in a range of fields, with image interpretation and time-series data analysis being common applications [48]. Such deep learning-based models can also be used for text classification, and they perform better in such tasks than existing machine learning-based models [49], [50], [51]. For example, Vaswani et al. [52] proposed Transformer, a new and simple deep learning network architecture based solely on an attention mechanism while excluding recurrence and convolutions. It achieved an overall Bilingual Evaluation Understudy (BLEU) score of 41.1, improving the existing best ensemble results by more than 1 and more than 0.7 for English–German and English–French translations, respectively. BERT also employs the encoder architecture of transformer to perform various tasks by understanding the contextual meaning of words. This differs from the mechanism used in generative pre-trained transformer (GPT), which is an NLP model that utilizes transformer's decoder architecture. When applied to supervised tasks (e.g., text classification), BERT demonstrated improved performance through fine-tuning. This approach involves pre-training the model on unstructured big data and then adapting it to a specific dataset [53], [54]. Despite the effectiveness of the model, to the best of the authors' knowledge, a few studies have used BERT to interpret traffic crash descriptions and classify the associated word meanings. Bareiss et al. [55] collected data from more than 9700 crash instances to train and validate a BERT-based text classification model for identifying pedal misapplication. When applied to the test dataset, the proposed BERT model reportedly demonstrated a classification accuracy of 95% for the four classes. The authors stated that the model can achieve reliable classification without the need for manual review in future. Meanwhile, Oliaee et al. [56] developed a methodology using BERT to classify five injury types, utilizing a large dataset of over 750,000 crash narrative reports. The reports were written from 2011 to 2016, and the classification accuracy for each year was approximately 83%. However, the result shows a deviation of more than 10% from three classes implying that model refinement and a different data preprocessing approach may be necessary.

Therefore, in the field of traffic safety research, a few studies have used NLP techniques to analyze traffic crash description data from crash reports. Although several recent studies have tried to classify meaningful words and identify traffic crash types, these are limited to identifying causal factors using the extracted words and have not attempted to extract new information that is not explained using the other crash report data elements. Therefore, this study proposes a practical BERT-based text classification methodology to achieve high performance in classification and aims to verify the utility of traffic crash description data by classifying narrative words and extracting new information from traffic crash reports to aid the understanding of crash scenes.

III. METHODS

A. OVERALL MODEL DEVELOPMENT PROCESS

The overall methodology for the classification and interpretation of unstructured traffic crash description data proposed in this study is shown in Figure 1 and consists of data preprocessing, model development, and model performance evaluation and validation. Unstructured traffic crash datasets, i.e., written descriptions of crashes, were collected and prepared for further analysis. The traffic crash description texts were first cleared using a word-embedding algorithm, including text preprocessing and NLP models, to correct typographical mistakes and heteronyms and remove sensitive wording referring to specific people or places. Second, a BERT-based text classification model was developed with detailed parameter settings. This model was applied to multiple subject wording groups representing traffic crash assailants, object maneuvers before the crash, and locations of damage on the objects involved in the crash. Finally, the classification and interpretation performance of the proposed model was evaluated against that of various other text classification models, including logistic regression, naïve Bayes, and SVM applied to the same datasets, and the performance of the BERT-based model was cross-validated.

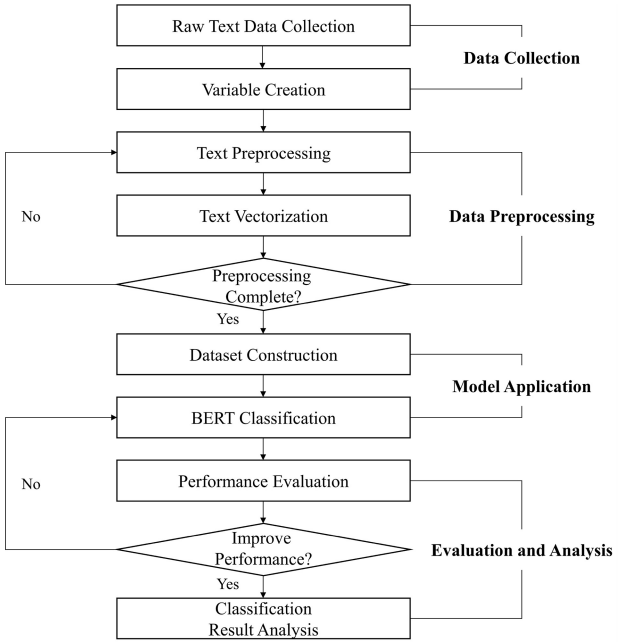


FIGURE 1. Overall procedure of model development.

B. DATA COLLECTION

A total of 2,427 records of traffic crashes that reported in Daegu, Republic of Korea during 2018 were obtained from the Korean Traffic Crash Analysis System. While a substantial amount of data plays a critical role in improving the prediction and analysis accuracy of text data processing, the available data is restricted to the 2018 annual record because it has not been disclosed from the Korea National Police Agency after 2018 due to privacy-related issues. Furthermore, the process of annotating traffic crash data necessitates a significant amount of time and effort. This annotation is an essential part of training a BERT model. Traffic crash data generally consist of multiple data elements, such as the crash date, time, type, and causal factors, as well as weather and pavement conditions, and most of the data elements are code-based inputs. Only one data column contained written descriptions of the crash situations, as recorded by police officers investigating the scene; for example, "While vehicle A was changing lanes at an intersection, it collided with the side of vehicle B, which was driving straight in the adjacent lane." The number of written descriptions varied based on the traffic crash record, the police officer who recorded it, and the complexity of the crash. In this study, such written descriptions were separated and extracted from the entire traffic crash data and used to develop the traffic crash description text classification and interpretation model. The extracted textual descriptions had an average of approximately 111 string units. The standard deviation, minimum number of strings, and maximum number of strings were 25, 35, and 252, whereas the first and third quartiles were 94 and 125, respectively. The deviation in traffic crash records with fewer than 94 strings was not large, whereas that of records with more than 125 strings was relatively large. As mentioned previously, sensitive wording and traffic crash records consisting of considerably short, incomplete, or indecipherable descriptions were excluded from the analysis.

Text Preprocessing: After the data had been collected, the text was preprocessed to convert it into a dataset for the text classification model. First, various words expressing the same meaning were modified into common words. This is because such synonyms may affect the classification performance by interrupting the extraction of sentence meaning. Misrepresented words, including typographical errors, are common, as officers generally write documents either manually or using computers; these were corrected.

Second, data cleansing was performed by identifying and deleting sensitive information, dates, and words identified according to the rules. The purpose of writing these words in traffic crash documents is not to describe crash situation but rather to preserve identifying information related to the vehicles involved; therefore, they are not crucial in text classification. Additionally, English words were excluded.

Finally, tokenization, which determines the basic units of inputs, was performed to quantify the text in the documents. A "token" is the minimum number of units that contain meaning, and various packages are provided for each language. One such package defines sentence pieces and has the advantage of tokenizing sentences without prior tokenization information. It can therefore be used for languages such as Chinese, Korean, or Japanese, where words are not clearly distinguished by spacing [54]. This study therefore used a sentence-piece as the tokenizer for text classification.

Text Vectorization: After text preprocessing, the tokens within a sentence were quantified in certain dimensions. Compared to previous methods for creating vectors with tokens in an embedding layer to perform NLP tasks (e.g., TF-IDF, Word2Vec, and FastText), in BERT, tokens are split into multiple vectors in multiple embedding layers, including token, segment, and position embedding layers [53]. The additional embedding layers enable BERT to gain a clearer understanding of the meaning of the contexts and therefore provide more accurate results.

C. MODEL DEVELOPMENT

Dataset Construction: The amount of data plays a crucial role in big data analysis, as low volumes may compromise the model's performance. This is because the model's effectiveness can improve with a substantial amount of data, and it requires both training and validation on different datasets. In this study, a K-fold cross-validation method was employed for training and validation of the model. This approach involves dividing the dataset into K subsets (or folds) of equal size. Among these sets of folds, K-1 folds are utilized for training, while the remaining folds are used for validation and testing. The process is subsequently repeated with a different set of folds until the classification

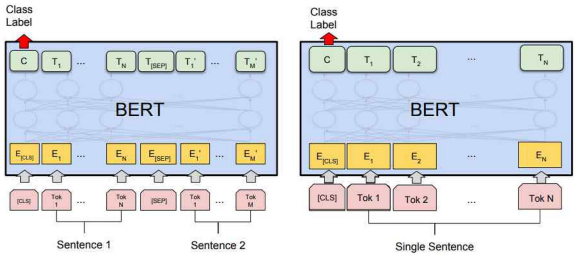


FIGURE 2. Framework of a BERT-based text classification model [53].

performance stabilizes. It is important to note that the same fold cannot be used for both training and testing in the same iteration, and the datasets used for training and testing are independent.

However, in imbalanced datasets, data cannot be equally distributed among folds based on the distribution of the classes, and K-fold cross-validation cannot ensure confidence. Stratified K-fold cross-validation has been suggested as a potential solution [57]. It involves generating data labels, identifying the amount of data in the entire dataset corresponding to each label, and constructing folds, including the same label ratio as that of the entire dataset. When using stratified K-fold cross-validation, the evaluation process involves calculating the average score after applying the selected metric to all K folds [58].

Bidirectional Encoder Representations from Transformer: BERT is an improved transformer-based model used for performing specific NLP tasks. Additionally, it compares every token in a sentence and based on the target task. The transformer consists of encoders for extracting the meaning of texts and decoders for performing the target task. The BERT model uses the same encoder structure of the basic transformer, as shown in Figure 2. A characteristic of BERT is that the learning process proceeds in both directions: from the beginning of the sentence to the end and from the end of the sentence to the beginning. This enables BERT to efficiently understand the meaning of sentences. Additionally, BERT models are generally pretrained with many non-task-specific texts. Through subsequent fine-tuning or relearning processes to rebalance the similarities between tokens and optimize the parameters in the models, their metric scores can be improved.

D. METRIC

For the classification of imbalanced data, dataset evaluation focuses on misclassification. The F1 score is one of the most representative metrics for multiclass classification and provides higher penalties for misclassification than other metrics. Because traffic crash datasets tend to be imbalanced, the F1 score was considered as the appropriate metric to evaluate the performance of each model in this study. The basic factors in the F1 score are defined as follows.

True Positive (TP): positive data are classified as positive

True Negative (TN): negative data are classified as negative

False Positive (FP): negative data are classified as positive

False Negative (FN): positive data are classified as negative

The formula for the F1 score includes precision and recall terms, which are described in Equation (1) and (2), where "precision" indicates the ratio of the positive data predicted to be positive by the model (TP) to the data predicted to be positive by the model (TP+FP), and "recall" is the ratio of the correct classification of positive data (TP) to actual positive data (TP+FN).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The F1 score is the harmonic mean of precision and recall in Equation (3). It is commonly used for classification metric because the F1 score is a more conservative measurement in classification than other alternatives including accuracy. In imbalanced datasets, the micro-averaged (Micro) F1 score considers the classification results of all classes simultaneously and can prevent such variability when the dataset is imbalanced. Therefore, this study uses the Micro F1 score.

$$F1 - Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

IV. RESULTS

A. TEXT CLASSIFICATION RESULTS

The target information for classification in the textual traffic crash descriptions comprised the maneuvers of the traffic crash-related entities (i.e., perpetrators and victims) immediately before the crash, detailed collision locations of the crash-related entities, and the identified perpetrator(s). This study used only vehicle-to-vehicle traffic crash records (which were more highly-represented than other types in the dataset) and created three analysis columns to extract the three aforementioned information items.

To define the text classification schema for the three targets, labels were first defined for each target, as listed in Table 1. For the first target (maneuvers of crash-related entities), nine labels were created, including straight, reverse, stop, left turn, right turn, lane change, and U-turn. For the collision locations, 13 labels were created to represent vehicle areas, including the front side, left side, right side, and rear side. The definition of each label in collision locations is shown in Figure 3. As an example, ① indicates the left front side of the vehicle and ⑧ indicates the right rear side of the vehicle. Finally, two labels indicating the perpetrator and victim were created for the third target (perpetrator).

TABLE 1. Text classification scheme labels.

Maneuver before the crash (Parameter I)		Collision location			
Label	Definition	Label	Definition	Label	Definition
0	Straight (normal Speed)	0	None	0	None
1	Straight (deceleration)	1	①	1	Vehicle A
2	Reverse	2	②	2	Vehicle B
3	Stop	3	③	-	-
4	Left turn	4	④	-	-
5	Right turn	5	⑤	-	-
6	Lane change	6	⑥	-	-
7	U-turn	7	⑦	-	-
8	Others	8	⑧	-	-
-	-	9	① and ②	-	-
-	-	10	③ and ⑤	-	-
-	-	11	④ and ⑥	-	-
-	-	12	⑦ and ⑧	-	-

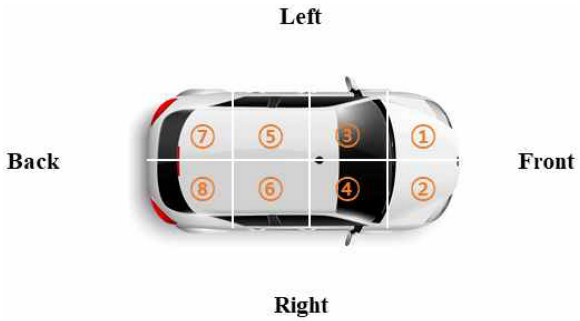


FIGURE 3. Collision location on vehicle.

Figure 4(a) summarizes the data regarding maneuvers of crash-related entities. Most cases involved driving straight ahead at a constant speed, followed by a lane change. The data is highly asymmetrical, which must be considered during classification. For example, the maneuvers of Vehicles A and B appear similar; however, "stop" was second-ranked for Vehicle B. Additionally, the data statistics on collision location (Figure 4(b)) show that for Vehicle A, frontal collisions were significantly more frequent than others, followed by rear collisions, indicating a severely unbalanced dataset. In the case of Vehicle B, rear-facing collisions (labeled "12") were the most frequent although frontal collisions are the most recorded in total ("1", "2", "9"). Therefore, despite an overall imbalanced dataset, the degree of imbalance in the data for Vehicle B was slightly lower than that for Vehicle A. Finally, for perpetrator classification, '1' was assigned to Vehicle A, and '2' was assigned to Vehicle B. When there was no perpetrator, or it was difficult to identify the perpetrator, all reports were classified as 0.

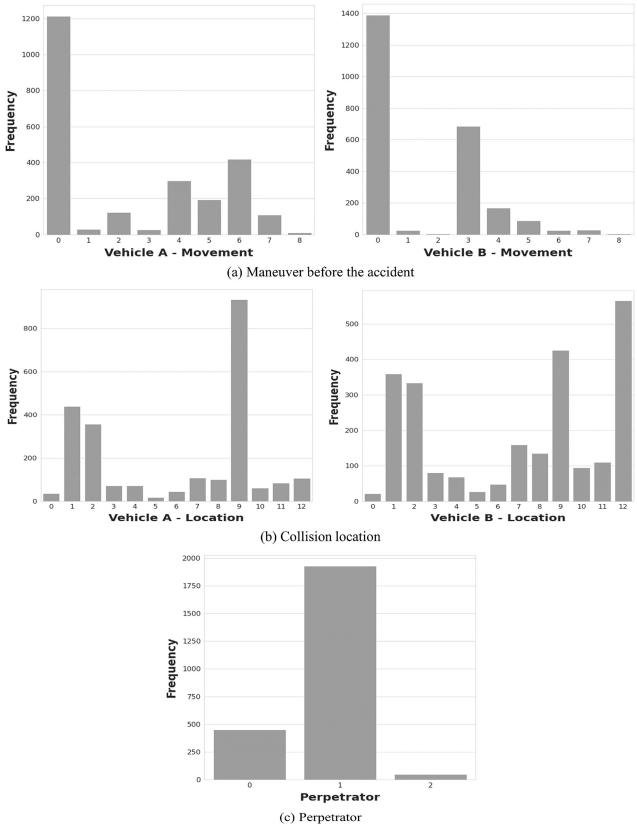


FIGURE 4. Data statistics of the classification of target information.

Figure 4(c) shows Vehicle A was the perpetrator in most cases and a noticeably large number of cases was observed where the perpetrator could not be identified.

B. DATA PREPROCESSING AND LEARNING

In the first step of text preprocessing, textual descriptions of collision location included various synonymous wordings for the vehicle bumper. This was presumed to be an issue caused by converting foreign language term(s) into Korean expressions. To eliminate this variability, all such cases were replaced with a representative word. Other words were subjected to the same process and were converted into simplified, uniform, and systematic wordings.

The data were simple to cleanse, as the vehicle identification numbers follow a consistent format in Korea (numbers-string-numbers), similar to the date (YYYY:MM) and time (HH:MM). Therefore, irrelevant or sensitive information entries were deleted from the entire dataset using their format characteristics.

The results obtained by applying the sentence-piece tokenizer used in BERT are shown in Figure 5. The average of number of tokens was approximately 64, and the standard deviation was approximately 14. The maximum, the minimum, the first quartile and the third quartile were 151, 23, 54, and 73 respectively.

Finally, text vectorization was performed to standardize the text, as follows: First, using existing learned tokens, each

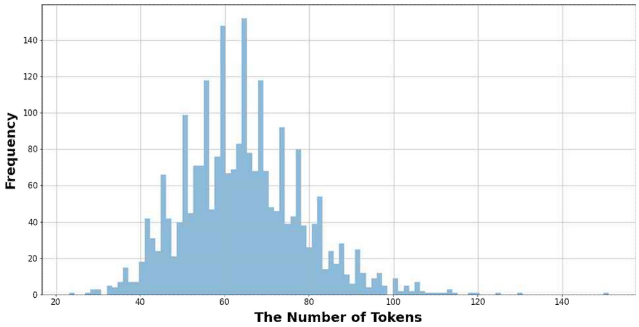


FIGURE 5. Frequency of tokens in the dataset.

sentence was divided into the smallest unit token, and the index and position of each token were embedded. Finally, a self-attention layer for learning was created, and the dataset to be applied to the model was constructed.

The following considerations were adopted during the development of the BERT model and implementation of the learning process. First, a stratified K-fold cross-validation algorithm was applied to facilitate the learning and evaluation of the unbalanced datasets. Five datasets were created for each column, resulting in 25 datasets. Subsequently, 25 BERT models were developed to correspond to each dataset, and the standardized vector was set to 90 dimensions. Classification was performed by designating the number of categories in each column, and learning was performed 20 times per model.

C. MODEL DEVELOPMENT AND PERFORMANCE EVALUATION

Traffic crash description text classification models were developed using BERT and combinations of various other state-of-the-art text classification algorithms, including logistic regression, naïve Bayes, SVM, CNN, and Bi-LSTM. Six algorithm combinations were used and compared: 1) TF-IDF + logistic regression, 2) TF-IDF + naïve Bayes, 3) TF-IDF + SVM, 4) Word2Vec + CNN, 5) FastText + Bi-LSTM, and 6) BERT.

In the case of BERT, the selection of parameters for model training is indeed a significant factor that can have a substantial impact on model performance. The key factors for model training include the maximum number of tokens per data point, the number of training iterations, and the learning rate based on the previous research [59].

These parameters play a crucial role in shaping the model's behavior during training and can greatly affect the model's overall performance. To determine those hyperparameters in the model, Sensitivity Analysis was conducted with 'Max Length,' which is a hyperparameter with substantial variation in text classification. A comparison of the results reveals that the 'Max Length' of 90 performs better in classification with F1-score. Therefore, hyperparameters in this study were selected as shown in Table 2.

These developed models were evaluated in terms of their classification performance on five types of traffic crash

TABLE 2. Model hyperparameters.

Type	Hyperparameters
Max Length	90
Batch Size	64
Warmup Ratio	0.1
Epoch	20
Learning Rate	0.00005

TABLE 3. Text classification performance of developed models.

Models	Micro F1 score				
	C1	C2	C3	C4	C5
TF-IDF + Logistic Regression	0.3016	0.4289	0.2225	0.1430	0.6815
TF-IDF + Naïve Bayes	0.1924	0.3070	0.1405	0.1133	0.5505
TF-IDF + SVM	0.2987	0.4330	0.2246	0.1401	0.6815
Word2Vec + CNN	0.8583	0.8913	0.6944	0.5835	0.8876
FastText + Bi-LSTM	0.9107	0.9531	0.7798	0.8271	0.9481
BERT	0.9778	0.9819	0.9489	0.9597	0.9465

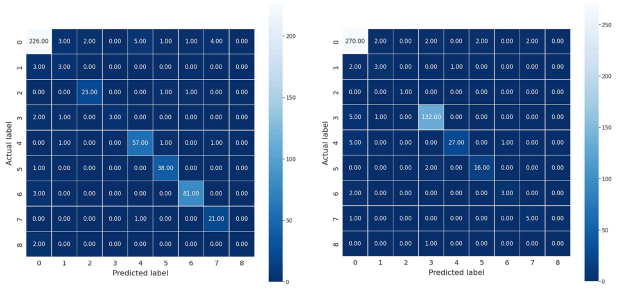
description text, as listed in Table 3. Here, C1 and C2 refer to maneuvers by the first and second vehicle, respectively, immediately before a vehicle-to-vehicle crash; C3 and C4 are the detailed collision location variables for the first and second vehicles, respectively; and C5 refers to a variable that determines the perpetrator.

Machine-learning algorithms, including logistic regression, naïve Bayes, and SVM, showed a lower classification performance than the other algorithms. The FastText tokenizer-based Bi-LSTM model performed better than the CNN-based algorithm and excelled against all models in the classification of perpetrator identification text. Finally, the proposed BERT-based model exhibited the highest performance for all variable classifications except perpetrator identification. Particularly, it exhibited significantly superior text classification performance in the collision location identification. Therefore, the BERT-based traffic crash situation text classification model effectively maintained its classification performance, even in the presence of many data labels.

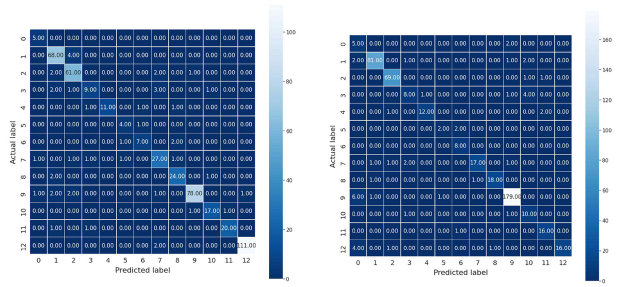
The cross-validation results for the BERT model used in this paper can be observed in Table 4. It is noticeable that there is a significant performance variation for C3 and C4, which have many labels. Additionally, there is notable performance variation in the category C5, which involves determining accident perpetrators. Identifying the reasons is crucial for a comprehensive evaluation of the model's performance and can help in optimizing the model for more consistent results in all categories.

TABLE 4. Text classification performance of 5 fold cross validation in bert.

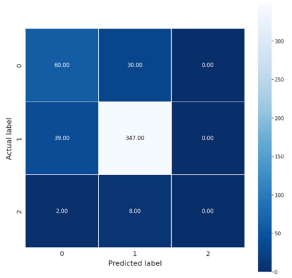
	C1	C2	C3	C4	C5
Fold 1	0.9300	0.9403	0.9074	0.9300	0.8374
Fold 2	0.9691	0.9815	0.9177	0.9095	0.9218
Fold 3	0.9918	0.9897	0.9340	0.9629	0.9794
Fold 4	0.9979	1.0000	0.9938	0.9959	0.9938
Fold 5	1.0000	0.9979	0.9918	1.0000	1.0000



(a) Maneuver before the accident (fold 1)



(b) Collision location (fold 1 and 2)



(c) Perpetrator

FIGURE 6. Text classification accuracy estimation results.

Figure 6 shows the confusion matrix, which illustrates the classification results for each classification item. Among them, Figure 6(a) shows the performance for the classification of "the maneuver before the crash"; Figure 6(b) shows that for "collision location", and Figure 6(c) indicates that for the "perpetrator" information.

The left side of Figure 6(a) indicates vehicle A, and the right side indicates Vehicle B. A high text classification accuracy was achieved for Vehicle A when the vehicle turned right (label = 5) or changed lanes (label = 6); the accuracy was also acceptable when the vehicle moved straight at a constant speed (label = 5), reversed (label = 0), and turned left (label = 4). However, misclassifications occurred in cases of deceleration (label = 1) and stopping (label = 3) while moving straight, owing to a lack of data samples. The text classification accuracy for Vehicle B was the highest when it moved straight at a constant speed (label = 0) or stopped (label = 3) compared to other labels; the classification performance was slightly low for decelerating, while going straight (label = 1), lane changing (label = 6), and U-turning (label = 7).

Figure 6(b) shows the performance of the text classification of the impact location for Vehicles A (left) and B (right), which was high overall. The highest accuracy was observed for collisions at the front of the vehicle (label = 1,2,7,9). A slightly poor performance was observed in cases where no collision occurred (label = 0) or when the collision was on the left side (label = 3) or right side (label = 6). In most cases, text classification was more accurate for Vehicle A than that for Vehicle B, particularly in front collisions. It was also good for the labels associated with collision points 1 and 2 (label = 9), left (label = 11), and rear (label = 12) locations.

Finally, as shown in Figure 6(c), when Vehicle A is the perpetrator (label = 1), the proposed approach exhibits a high classification performance compared to the other labels. When Vehicle B was the perpetrator (label = 2), many misclassifications were observed. However, even when the data could not identify the perpetrator (label = 0), many patterns were obtained for distinguishing Vehicle A as the perpetrator.

V. DISCUSSION ON PERFORMANCE DEGRADATION

The BERT-based traffic crash description text classification model proposed in this study showed superior performance for the classification of text variables describing traffic crash situations compared to regression and machine learning-based algorithms. However, it could not effectively identify certain labels, which translated to low performance for specific cases. Labels with a low text classification performance for each of the three text variables of movement (immediately before the crash, collision location, and perpetrator identification) were examined. There were cases in which it was difficult to classify keywords with relatively few cases (such as overtaking and centerline violations). In the detailed collision location variable, many cases existed where the model identified no collisions, despite the indication that the collision occurred. This attributed to insufficient learning owing to a lack of samples. Relatively few cases of non-collisions in traffic crash descriptions are present, and additional datasets are needed to reduce such misunderstandings. For the perpetrator identification variable, the perpetrator information derived by analyzing the external meaning of the sentences presented in the traffic crash descriptions and that derived by analyzing the actual legal responsibility may differ. In such cases, the text classification

performance can be degraded. Therefore, to improve the performance of the BERT-based traffic crash description text classification model, additional training data (i.e., traffic crash description records) and efforts to standardize the sentence structures of traffic crash descriptions are required.

VI. CONCLUSION

This study developed a methodology for classifying important words describing traffic crash situations into standardized data using the BERT model (an NLP technique) for unstructured textual data describing traffic crash situations. The text classification performance of the proposed model was compared with that of various state-of-the-art NLP algorithms, including regression- and machine-learning-based algorithms, which have previously been applied to develop text classification models for traffic crash descriptions. Among the NLP algorithms, the BERT-based model exhibited the highest performance in the interpretation of three traffic crash elements (maneuvers before the crash, collision locations on vehicles, and perpetrator). Furthermore, the BERT-based model showed a superior text classification performance for each label of these three elements that can be extracted from the traffic crash description texts (such labels, for example, numbers corresponding to vehicle regions, are considered important for reconstructing crash scenes) compared to other methods. In particular, the text classification accuracy of the Bert-based model is higher than 95% for most text labels, indicating that it is well-trained and is an appropriate tool for the automatic analysis of crash descriptions.

This study provides a practical method for analyzing and interpreting unstructured textual traffic crash descriptions among traffic crash data elements by converting them into standardized data that can be analyzed using NLP. The text classification methodology proposed in this study is expected to increase the number of application cases of this under-utilized data in traffic safety research, enabling the in-depth analyses of the causes of traffic crashes. This will ultimately reduce the number of traffic crashes and improve traffic safety. Regarding the scientific contribution of this study, the text classification algorithm proposed herein can analyze the characteristics of a dataset with a low classification accuracy in the initial text classification result and improve that result by applying a learning system that utilizes additional rule-based text preprocessing.

However, the text classification model proposed needs improvement through the addition of extra training and test samples and the usage of uniform crash descriptions. Due to the insufficient quantity of traffic crash record data used in this study, achieving high text classification performance was not possible. As a result, future research should aim to rectify this issue. To attain an F1 score approaching 1.0, indicating better accuracy for text classification compared to this study, more traffic crash records are needed. Moreover, developing models that take into account an ample supply of uniform traffic crash description texts in terms of both quality and quantity is necessary. Developing a methodology for reconstructing traffic crash situations through the fusion of data extracted from crash descriptions and other structured data elements in traffic crash records would greatly enhance our comprehension of traffic crashes.

REFERENCES

[1] K. Bhatt, N. Gore, J. Shah, and S. Arkatkar, "Drivers' dilemma at high-speed unsignalized intersections," *Transp. Res. Rec.*, Jun. 2023, Art. no. 03611981231178813.

[2] M. M. Hossain, H. Zhou, and S. Das, "Data mining approach to explore emergency vehicle crash patterns: A comparative study of crash severity in emergency and non-emergency response modes," *Accid. Anal. Prevent.*, vol. 191, Oct. 2023, Art. no. 107217.

[3] H. Nassereddine, K. R. Santiago-Chaparro, and D. A. Noyce, "Evaluating right-turn flashing yellow arrow for vehicle–pedestrian interactions using a non-probabilistic regression approach," *Transp. Res. Rec.*, Jun. 2023, Art. no. 03611981231173645.

[4] O. Olufowobi, J. Ivan, S. Zhao, K. Wang, and N. Eluru, "Application of realistic artificial data for testing various crash safety analyses: A case study for rural two-lane undivided highways," *Transp. Res. Rec.*, Jun. 2023, Art. no. 03611981231175901.

[5] P. Puthan, N. Lubbe, J. Shaikh, B. Sui, and J. Davidsson, "Defining crash configurations for powered two-wheelers: Comparing ISO 13232 to recent in-depth crash data from Germany, India and China," *Accid. Anal. Prevent.*, vol. 151, Mar. 2021, Art. no. 105957.

[6] X. Wang, Y. Peng, S. Yi, H. Wang, and W. Yu, "Risky behaviors, psychological failures and kinematics in vehicle-to-powered two-wheeler accidents: Results from in-depth Chinese crash data," *Accid. Anal. Prevent.*, vol. 156, Jun. 2021, Art. no. 106150.

[7] P. Nitsche, P. Thomas, R. Stuetz, and R. Welsh, "Pre-crash scenarios at road junctions: A clustering method for car crash data," *Accid. Anal. Prevent.*, vol. 107, Oct. 2017, pp. 137–151.

[8] F. Gidion, J. Carroll, and N. Lubbe, "Motorcyclist injuries: Analysis of German in-depth crash data to identify priorities for injury assessment and prevention," *Accid. Anal. Prevent.*, vol. 163, Dec. 2021, Art. no. 106463.

[9] M. Schmidt, S. A. J. Schmidt, J. L. Sandegaard, V. Ehrenstein, L. Pedersen, and H. T. Sørensen, "The danish national patient registry: A review of content, data quality, and research potential," *Clin. Epidemiol.*, vol. 7, pp. 449–490, Nov. 2015.

[10] S. Kim, J. Lee, and Y. Youn, "A study on the construction of the database structure for the Korea in-depth accident study," *Trans. Korean Soc. Automot. Eng.*, vol. 22, no. 2, pp. 29–36, Mar. 2014.

[11] D. Otte, M. Jänsch, and C. Haasper, "Injury protection and accident causation parameters for vulnerable road users based on German in-depth accident study GIDAS," *Accid. Anal. Prevent.*, vol. 44, no. 1, pp. 149–153, Jan. 2012.

[12] R. Ganguli, P. Miller, and R. Pothina, "Effectiveness of natural language processing based machine learning in analyzing incident narratives at a mine," *Minerals*, vol. 11, no. 7, p. 776, Jul. 2021.

[13] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," *Comput. Ind.*, vol. 78, pp. 80–95, May 2016.

[14] S. Park, S. Park, H. Jeong, I. Yun, and J. J. So, "Scenario-mining for level 4 automated vehicle safety assessment from real accident situations in urban areas using a natural language process," *Sensors*, vol. 21, no. 20, p. 6929, Oct. 2021.

[15] J. J. So, I. Park, J. Wee, S. Park, and I. Yun, "Generating traffic safety test scenarios for automated vehicles using a big data technique," *KSCE J. Civ. Eng.*, vol. 23, pp. 2702–2712, Mar. 2019.

[16] K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*. New Delhi, India: Springer, 2020, pp. 603–649.

[17] E. Kumar, *Natural Language Processing*. New Delhi, India: I K Int. Pvt. Ltd., 2011.

[18] J. Eisenstein, *Introduction to Natural Language Processing*. Cambridge, MA, USA: MIT press, 2019.

[19] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how bert works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, Jan. 2021.

[20] P. Chakraborty, Y. O. Adu-Gyamfi, S. Poddar, V. Ahsani, A. Sharma, and S. Sarkar, "Traffic congestion detection from camera images using deep convolution neural networks," *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 222–231, 2018.
[21] H. H. Chen, Y. B. Lin, L. H. Yeh, H. J. Cho, and Y. J. Wu, "Prediction of queue dissipation time for mixed traffic flows with deep learning," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 267–277, 2022.
[22] A. J. Huang and S. Agarwal, "Physics-informed deep learning for traffic state estimation: Illustrations with LWR and CTM models," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 503–518, 2022.
[23] V. Papathanasopoulou, I. Spyropoulou, H. Perakis, V. Gikas, and E. Andrikopoulou, "A data-driven model for pedestrian behavior classification and trajectory prediction," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 328–339, 2022.
[24] Z. Wang et al., "Classification of automated lane-change styles by modeling and analyzing truck driver behavior: A driving simulator study," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 772–785, 2022.
[25] A. Gholamhosseinian and J. Seitz, "Vehicle classification in intelligent transport systems: An overview, methods and software perspective," *IEEE Open J. Intell. Transp. Syst.*, vol. 2, pp. 173–194, 2021.
[26] K. Yang, C. Al Haddad, G. Yannis, and C. Antoniou, "Classification and evaluation of driving behavior safety levels: A driving simulation study," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 111–125, 2022.
[27] S. Zhang and M. Abdel-Aty, "Real-time pedestrian conflict prediction model at the signal cycle level using machine learning models," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 176–186, 2022.
[28] R. Alms, A. Noulis, E. Mintsis, L. Lücken, and P. Wagner, "Reinforcement learning-based traffic control: Mitigating the adverse impacts of control transitions," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 187–198, 2022.
[29] M. Emu, F. B. Kamal, S. Choudhury, and Q. A. Rahman, "Fatality prediction for motor vehicle collisions: Mining big data using deep learning and ensemble methods," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 199–209, 2022.
[30] R. Valiente, B. Toghi, R. Pedarsani, and Y. P. Fallah, "Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 397–410, 2022.
[31] M. Atif, A. Ceccarelli, T. Zoppi, M. Gharib, and A. Bondavalli, "Robust traffic sign recognition against camera failures," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 709–722, 2022.
[32] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR," *IEEE Trans. Intell. Veh.*, vol. 8, no. 8, pp. 4069–4080, Aug. 2023.
[33] Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," *Accid. Anal. Prevent.*, vol. 108, pp. 122–130, Nov. 2017.
[34] L. Gao and H. Wu, "Verb-based text mining of road crash report," in *Proc. TRB 92nd Annu. Meet.*, 2013, p. 12.
[35] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accid. Anal. Prevent.*, vol. 88, pp. 37–51, Mar. 2016.
[36] X. Wan, M. C. Lucic, H. Ghazzai, and Y. Massoud, "Empowering real-time traffic reporting systems with nlp-processed social media data," *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 159–175, 2020.
[37] T. Fontes, F. Murçós, E. Carneiro, J. Ribeiro, and R. J. Rossetti, "Leveraging social media as a source of mobility intelligence: An NLP-based approach," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 663–681, 2023.
[38] A. Salas, P. Georgakis, and Y. Petalas, "Incident detection using data from social media," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, 2017, pp. 751–755.
[39] M. Noaeen and B. H. Far, "The efficacy of using social media data for designing traffic management systems," in *Proc. 4th Int. Workshop Crowd-Based Requir. Eng. (CrowdRE)*, 2020, pp. 11–17.
[40] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (TF-IDF) heuristic (and variations motivated by this explanation)," *Int. J. Gener. Syst.*, vol. 46, no. 1, pp. 27–36, 2017.
[41] K. W. Church, "Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017.
[42] P. Rodríguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image Vis. Comput.*, vol. 75, pp. 21–31, Jul. 2018.
[43] S. Selva Birunda and R. Kanniga Devi, "A review on word embedding techniques for text classification," in *Proc. ICIDCA*, 2020, pp. 267–281.
[44] L. Ge and T.-S. Moh, "Improving text classification with word embedding," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2017, pp. 1796–1805.
[45] S. M. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," 2017, *arXiv:1711.08609*.
[46] J. Gu et al., "Recent advances in convolutional neural networks," in *Proc. Pattern Recognit.*, vol. 77, 2018, pp. 354–377.
[47] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*.
[48] R. Valiente et al., "Robust perception and visual understanding of traffic signs in the wild," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 611–625, 2023.
[49] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning—Based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2021.
[50] C. N. Kamath, S. S. Bukhari, and A. Dengel, "Comparative study between traditional machine learning and deep learning approaches for text classification," in *Proc. ACM Symp. Doc. Eng.*, 2018, pp. 1–11.
[51] C. Suneera and J. Prakash, "Performance analysis of machine learning and deep learning models for text classification," in *Proc. IEEE 17th India Council Int. Conf. (INDICON)*, 2020, pp. 1–6.
[52] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[54] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," 2018, *arXiv:1808.06226*.
[55] M. Bareiss, C. Smith, and H. C. Gabler, "Finding and understanding pedal misapplication crashes using a deep learning natural language model," *Traffic Inj. Prevent.*, vol. 22, no. Sup1, pp. S169–S172, 2021.
[56] A. H. Oliaee, S. Das, J. Liu, and M. A. Rahman. "Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types," *Nat. Lang. Process. J.*, vol. 3, Art. no. 100007, Jun. 2023.
[57] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves. "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.
[58] Y.-Q. Liu, C. Wang, and L. Zhang, "Decision tree based predictive models for breast cancer survivability on imbalanced data," in *Proc. 3rd Int. Conf. Bioinf. Biomed. Eng.*, 2009, pp. 1–4.
[59] Y. S. Lee, H. R. Jang, S. B. Oh, Y. I. Yoon, and T. W. Um, "Prediction of content success and cloud-resource management in internet-of-media-things environments," *Electronics*, vol. 11, no. 8, p. 1284, 2022.

YOUNGHOON SEO received the B.S. degree from the University of California at San Diego, La Jolla, in 2017, and the M.E. degree from Sungkyunkwan University in 2019. He is currently a Senior Researcher with the Intelligent Transportation Laboratory, Advanced Institute of Convergence Technology. His research interests include machine learning, text mining, and autonomous driving.

JIHYEOK PARK received the B.S. degree and the M.S. degree in transportation engineering from Ajou University. During his master's course, he conducted research on the development of legal regulations and acceptability indicators for Autonomous driving cars. His research interests include determining whether autonomous vehicles comply with legal regulations through simulation and determining driving safety.

GYUNGTAEK OH received the B.S. and M.S. degrees in transportation system engineering from Ajou University. He is working for Autonomous Vehicles with the Advanced Institute of Convergence Technology. As his engineering master's thesis, he wrote "Optimization of Automated Driving Parameters Based on Road Hierarchy and Characteristic". During his master's studies, he worked on research projects in the fields of autonomous driving and smart cities.

HYUNGJOO KIM received the B.S. degree from Ajou University in 2010, the M.S. degree from Seoul National University in 2012, and the Ph.D. degree from the Korea Advanced Institute of Science and Technology in 2019. He is currently the Director of the Intelligence Transportation System Laboratory, Advanced Institute of Convergence Technology. His research interests are traffic flow analysis and modeling, intelligent transportation systems, and autonomous vehicle.

JIA HU (Member, IEEE) works as a ZhongTe Distinguished Chair of Cooperative Automation with the College of Transportation Engineering, Tongji University. Before joining Tongji University, he was a Research Associate with FHWA, USA. He is an Associate Editor of the *Journal of Transportation Engineering* (ASCE) and the IEEE OPEN JOURNAL IN INTELLIGENT TRANSPORTATION SYSTEMS and an Assistant Editor of the *Journal of Intelligent Transportation Systems*. He is a member of TRB (a Division of the National Academies) Committees and the ASCE Transportation and Development Institute Committees.

JAEHYUN (JASON) SO (Member, IEEE) is an Assistant Professor with the Department of Transportation System Engineering, Ajou University, Suwon, Republic of Korea, where he is operating the mobility operation design, validation, and enhancement laboratory "move lab" and leading various research projects in ITS/C-ITS, automated vehicle, smart mobility, and mobility data. Before joining Ajou University, he worked four years with the Center of Smart City and Transport, Korea Transport Institute from October 2016 to August 2020 and led many national smart city projects and studies. Before back in South Korea, he worked a total of four years with Florida Atlantic University, USA, from August 2013 to October 2014 and the Technical University of Munich, Germany, from November 2014 to September 2016 as a Research Scientist and a Lecturer.