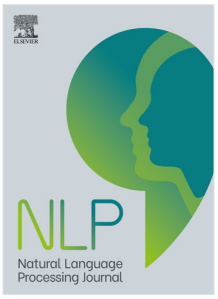


Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types

Amir Hossein Oliaee, Subasish Das, Jinli Liu, M. Ashifur Rahman

PII: S2949-7191(23)00004-3
DOI: https://doi.org/10.1016/j.nlp.2023.100007
Reference: NLP 100007

To appear in: *Natural Language Processing Journal*

Received date : 27 November 2022
Revised date : 13 April 2023
Accepted date : 14 April 2023

Please cite this article as: A.H. Oliaee, S. Das, J. Liu et al., Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. *Natural Language Processing Journal* (2023), doi: https://doi.org/10.1016/j.nlp.2023.100007.

Using Bidirectional Encoder Representations from Transformers (BERT) to Classify Traffic Crash Severity Types

Amir Hossein Oliaee, Ph.D. Student
Department of Architecture, Texas A&M University
Langford Architecture Building 3137, College Station, TX 77840
E-mail: oliaee@tamu.edu
ORCID: 0000-0002-0574-2690

Subasish Das, Ph.D.
(Corresponding author)
Assistant Professor, Ingram School of Engineering
Texas State University
601 University Drive, San Marcos, Texas 78666
E-mail: subasish@txstate.edu
ORCID ID: 0000-0002-1671-2753

Jinli Liu, Ph.D. Student
College of Liberal Arts, Texas State University
601 University Drive, San Marcos, Texas 78666
E-mail: j_l848@txstate.edu
ORCID: 0000-0002-6152-8808

M. Ashifur Rahman, Ph.D.
Research Associate, University of Louisiana at Lafayette
E-mail: ashifur@louisiana.edu
ORCID: 0000-0001-6940-1599

Abstract

Traffic crashes are a critical safety concern. Many studies have attempted to improve traffic safety by performing a wide range of studies on safety topics with the application of diverse statistical and machine learning models. The data elements contained in police-reported crash narrative information are not routinely analyzed with coded and structured crash data. In the recent years, unstructured textual contents in traffic crash narratives have been investigated by many researchers. However, most of these studies are basic text mining applications and often the dataset is limited in size. This study applied an advanced language model Bidirectional Encoder Representations from Transformers (BERT) to classify traffic injury types by using a dataset of over 750,000 unique crash narrative reports. The models have an 84.2% ±0.5 predictive accuracy and an Area Under the receiver operating Curve (AUC) of 0.93 ±0.06 per class. Overall, the findings can assist safety engineers and analysts in determining the causes of a crash. The classification of crash injury types using a language model like BERT is a valuable tool for identifying additional factors that contribute to crashes, which can identify new areas for safety countermeasures and support the development of new safety strategies.

Keywords: crashes, safety, severity, text mining, BERT, language model.

1. Introduction

In the United States, it is common for crash reports to include a narrative that includes a written summary of the crash by a police officer. Crash narratives contain useful information that can aid in understanding the circumstances surrounding a crash at a specific roadway location. However, because there are hundreds of thousands of reports, the crash report narratives contain unstructured textual information that is difficult to extract and use in analyses. The proliferation of digital textual archives in the domain of transportation safety necessitates the development of efficient methods for extracting information from textual data sources.

NLP tools can facilitate accurate crash severity prediction which has a significant societal impact. In addition to identifying crash severity for incidents with unknown severity, it can help transportation safety planners, hospitals, and insurance companies in predicting future crash costs, providing timely medical care, and determining customer premiums (Iranitalab and Khattak, 2017). By improving safety measures and reducing the impact of crashes, it can have far-reaching benefits.

Many studies have been conducted in recent years to investigate the utility of unstructured crash narrative reports in uncovering hidden insights. To provide frequency or correlation-based findings, most of these studies used basic natural language processing (NLP) tools such as word frequency analysis (Das et al., 2021), word cloud (Krause and Busch 2019), unigram or bi-gram analysis (Zhang 2020), and text network analysis (Das et al., 2021). Some studies (Kwayu et al., 2021; Wali et al., 2021; Bareiss et al., 2021) used advanced NLP tools to identify information or patterns that are typically inaccessible in conventional structured crash records. As far as the researchers of this study are aware, there is only one previous study that used advanced language models such as Bidirectional Encoder Representations from Transformers (BERT) to solve transportation safety problems, and that study is Bareiss et al. (2021). In their study, a system that made use of the BERT natural language comprehension model was developed to detect pedal misapplication (PM) crashes based on crash narrative.

In this study, the proposed fine-tuned BERT-based model can handle complex sentence structures like clauses and conjunctive sentences with ease. In the context of this research, fine-tuning refers to the training of the model using the pre-trained weights from the original (BERT) implementation in the initial step of the process. Random initial weights were used for the appended portion of the model. Given the noticeable impact of the BERT on subsequent NLP research by being able to address the problem of text ambiguity, insight into its capability to classify crash narratives without any post-prediction processing could provide a valuable baseline for future research. The results also provide a point of comparison for future models, fine-tuning methods, and effects of preprocessing and post-processing practices on crash narrative classification. The proposed method was tested using crash data from Louisiana from 2011 to 2016. The results show that the extracted information is useful not only for crash classification but also for understanding crash causes.

The researchers acknowledge the existence of more recently developed NLP models that have outperformed BERT on various tasks or proved to be a lighter implementation requiring less computational resources: Transformer-XL (Dai et al., 2019); XL-net (Yang et al., 2020); RoBERTa (Liu et al., 2019); DistilBERT (Sanh et al., 2020); ALBERT (Lan et al., 2020); MobileBERT (Sun et al., 2020); ELECTRA (Clark et al., 2020); ConvBERT (Jiang et al., 2020); DeBERTa (He et al., 2020); BigBird (Zaheer et al., 2020). However, BERT serves as a building block of most these models and remains an important point of comparison for current NLP tasks.

Based on the recent advancements of NLP models and their broad applicability for traffic safety tasks, researchers deemed it necessary to assess the performance of this significant model and share the findings to facilitate implementation of models and techniques that are beyond the scope of this research.

This study is unique form several aspects. The study is the first to apply fine-tuned BERT-based model on a very large crash narrative dataset collected statewide of approximately 750,000 crashes spread across six years for comprehensive crash severity typing. The substantially large sample size of crashes sets this study apart from other studies. This study is also unique in terms of its practicality in its classification approach. Unlike previous studies that evaluated BERT as a single binary classification model, this study models multi-label classification based on the law-enforcement agency's 5-label standard classification. Consequently, this study opens up new avenues for future research on incorporating more advanced fine-tuning mechanisms and expands the practicality of accurately classifying transportation safety-related issues.

2. Literature Review

In the recent years, many safety studies have incorporated different NLP tools to explore insights from traffic safety-related textual contents.

A few studies (Gao et al., 2013; Goh and Ubeynarayana, 2017) attempted to explore crash narrative reports to classify crash types. Gao and Wu (2013) created a verb-based text mining method by exploring around 1,000 traffic crash narrative records from Missouri. The results showed that the extracted information is beneficial not only for crash classifications but also for identifying causal patterns. Goh and Ubeynarayana (2017) examined the usefulness of different text mining classification techniques for classifying one thousand publicly available construction crash narratives. They found that linear SVM with unigram tokenization and radial basis function (RBF) SVM with unigram tokenization were the most effective classifiers.

Several recent studies applied crash narrative mining on different transportation safety problems: pedestrian crash (Das et al., 2020), intersection crashes (Kwayu et al., 2020), culvert-related crashes (Chawla et al., 2021), motorcycle crash data (Das et al., 2021), work zone crash data (Sayed et al., 2021), tree and utility pole crashes (Das et al., 2021), bicycle crashes (Lopez et al 2022), and speeding-related crashes (Fitzpatrick et al 2017). Das et al. (2020) developed a mechanism for classifying pedestrian-related crash types from unstructured text using different machine-learning models. It was determined that the XGBoost model was the best classifier, indicating that this machine-learning technique could categorize pedestrian crash types with a maximum accuracy of up to 72 percent for test data. Kwayu et al. (2020) utilized semantic analysis to determine the most probable crash scenario at signal-controlled junctions for each hazardous activity with respect to driver maneuver. Chawla et al. (2021) evaluated the in-service safety performance of roadside culverts and the prospective effects of applying various safety interventions to reduce the severity of crashes involving culverts. Das et al. (2021) analyzed the unstructured linguistic content of the data from the motorcycle crash causation study (MCCS) through the application of various NLP methods (e.g., text mining, topic modeling). The narratives of fatal and nonfatal collisions were clustered individually to acquire insight into injury severity. The research conducted by Sayed et al. (2021) on work zone safety management mainly relied on the quality of work zone crash data. Based on the crash narrative reports, an NLP classifier was developed. Das et al. (2021) explored the fact that despite significant advancements in vehicle safety, roadway design, and operations, there are still an excessive number of traffic collisions that

result in casualties and significant productivity losses. This study therefore utilized text mining and interpretable machine learning (IML) approaches to assess all tree and utility pole crashes (with known crash narratives) that happened in Louisiana between 2010 and 2017. Among the employed modeling strategies, the eXtreme gradient boosting (XGBoost) model performed better. Lopez et al. (2022) investigated the data quality of bicycle crash narratives in police reports. Police records were compared with the Pedestrian and Bicycle Crash Analysis Tool (PBCAT) to determine how much information was gathered and when more information was likely to be captured. Fitzpatrick et al. (2017) examined the speeding-related crash classification by developing a set of logistic regression models derived from the existing speeding-related collision typologies.

Several studies (Bunn et al., 2008; Kim et al., 2021; Møller et al., 2021) explored agriculture, farm-related crashes and hospital records to identify important insights. Bunn et al. (2008) performed narrative text analysis on 69 Kentucky Fatality Assessment and Control Evaluation (FACE) agricultural tractor fatality reports. Kim et al. (2021) examined structured or coded data fields from a crash report as the basis for identifying crashes involving various vehicle categories, such as agricultural equipment. Using Texas and Louisiana crash accounts from various time periods, several data forms and document classification techniques were investigated. Møller et al. (2021) conducted an in-depth examination of the free-text fields in hospital records to identify elements linked with each crash using four categories: cyclist, road, bicycle, and other party. The authors utilized the chi-square test to analyze the distribution of characteristics across hospital and police collision reports.

Advanced NLP models such as topic models (Kwayu et al., 2021; Wali et al., 2021) and language models such as BERT (Bareiss et al., 2021) have been utilized in several transportation safety studies. Kwayu et al. (2021) used crash narratives in conjunction with accident metadata to determine the prevalence and co-occurrence of crash-related themes. The structural topic modeling (STM) and network topology analysis were utilized to produce and investigate the prevalence and interaction of themes from the crash narratives, which were primarily classified as pre-crash events, crash locations, and involved parties. Wali et al. (2021) took advantage of the rapid improvements in more sophisticated qualitative analysis techniques for detecting thematic concepts in unstructured narrative materials on crashes. The research found a statistically significant correlation between the presence of a crash narrative and the severity of the trespasser's injuries (coded as minor, major, and fatal injury). Bareiss et al. (2021) created a system that utilized the BERT natural language comprehension model to identify pedal misapplication (PM) crashes from crash narratives and validate the system's accuracy. After training, the language model was applied to a test dataset of 8,668 North Carolina and National Motor Vehicle Crash Causation Survey (NMVCCS) cases to evaluate the occurrence of pedal misapplication.

Some studies used news media (Das, 2021) and reports (Zhang et al., 2021) to explore safety-related textual contents. Das (2021) examined the ideas that explain why road collisions are a significant public health hazard. This study collected fatal crash stories from online English daily newspapers using a Google News Alert. The findings of this study indicated that online news coverage of road deaths tends to differ between news organizations. Zhang et al. (2021) utilized a series of data-mining and sequential deep learning approaches on National Transportation Safety Board (NTSB) accident investigation records in support of the prediction of adverse occurrences. The development of classification models for passenger air carriers as well as a prototype for an interactive query interface enabling end-users to evaluate various situations, including entire or partial event sequences or narratives, and obtain predictions for unfavorable events.

The literature review reveals that most of the past studies applied basic NLP tools to identify patterns or classify injury types or other relevant classification related tasks. Bareiss et al. (2021) applied BERT-based model on a dataset of 11,637 samples. However, even though they fine-tuned the BERT-based model for a multi-label classification task, they evaluated the model as a single binary classification model and based their conclusion on Area under Receiver Operating Characteristics Curve (AUROC), which does not provide an insight into the model's multi-class labelling capability and challenges that come with this type of task. According to the knowledge of the researchers of this study, this is the first application of a BERT-based model on a dataset of approximately 750,000 crash records that provides a comprehensive assessment of fine-tuned BERT as a crash narrative typing tool. The methodology and findings can provide important guidance in applying this approach in other critical transportation safety issues.

3. Data Preparation

For this study, the research team collected crash narrative data from Louisiana. Six years (2011-2016) of crash data (with narrative reports) were collected by removing the personal identification information. After compiling data for six years, the research team used a threshold of minimum 15 words in each narrative report to remove the crashes with insufficient information. A total of 740,736 crash reports were finally retrieved with crash narrative data.

According to Highway Safety Manual, crash severity can be used for establishing the level of injury caused by a crash cost. KABCO scale is frequently used by law enforcement for classifying injuries. The definition of KABCO is as follows:

- K- Fatal injury
- A - Severe injury
- B - Moderate injury
- C - Complaint
- O - No injury

Table 1 presents the frequencies of these reports by injury type and year. The reason behind using 2011-2016 data is the availability of these structural textual contents being limited till 2016.

Table 1
Counts of crash narrative reports by injury type and year.

Year	Injury Type					Yearly Total
	Fatal (K)	Severe (A)	Moderate (B)	Complaint (C)	No Injury (O)	
2011	278	1,038	6,995	22,954	82,965	114,230
2012	238	1,009	7,353	24,156	85,440	118,196
2013	261	1,041	7,287	23,742	87,309	119,640
2014	281	985	7,326	24,853	88,320	121,765
2015	302	1,027	7,816	27,115	95,078	131,338
2016	291	1,031	7,965	28,294	97,986	135,567
Severity Total	1,651	6,131	44,742	151,114	537,098	740,736

*Note: *The counts do not represent actual crash counts by severity. These numbers indicate count of crash reports by injury type with available crash narrative report.*

The quality of the narrative data greatly impacts the model's ability to infer the level of severity of the crash. To limit the irrelevant information and maximize the inference capability of the model, a series of preprocessing steps were first taken. These steps were uniformly applied to the dataset in its entirety and did not include any data augmentation steps.

The preprocessing started with replacement of all the non alphabetic characters (including numbers and newline characters) with whitespace characters. An exception was made for the period character to maintain the sentence separation. Then, the duplicated whitespace characters, be they pre-existing or caused by the replacement process, were removed. The resulting narrative data was much shorter in length. After an overview of the preprocessing results, the data was tokenized. The initial 510 token for each narrative was only used with addition of a [CLS] prefix token to mark the start and a [SEP] suffix token to mark the end of the sequence. The dataset was then randomly divided into training (70%), test (20%), and validation (10%) subsets. The validation data was not used in the model's development process.

4. Methodology

This section provides a comprehensive summary of the proposed crash narrative classification based on BERT architecture.

4.1 BERT-based Model

One of the most influential language models in recent years is BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformers. The BERT model consists of encoder blocks. However, the encoder block was previously introduced as part of transformer architecture developed by Vaswani et al. (2017). The bidirectional nature of this model helps the model better interpret words in their context resulting in improved accuracy. However, the BERT architecture consists of millions of parameters that require substantial data and computation to effectively adjust through training. Given the general applicability of this model and the variety of data that the available pretrained BERT models ($BERT_{BASE}$ and $BERT_{LARGE}$) have been exposed to, these pretrained models could be applied with relative ease.

The pretraining process consists of two tasks. The Masked Language Modeling (MLM), also referred to as the Cloze task (Taylor, 1953), is done through random masking of the input tokens and prediction of the masked words. The Next Sentence Prediction (NSP) task trains the model about sequence relationships. This is done by pre-training the model for binarized next sentence prediction.

4.2 Word Embeddings

The text classification models, much like other mathematical models, perform a task based on numerical input. The process of converting raw text to a numerical vector input, also known as encoding, varies based on the model's architecture and the intended task. However, some of the most notable text encoding methods include Word2Vec (Mikolov et al.), Global Vectors (GloVe) (Pennington et al., 2014), and FastText (Joulin et al., 2017). Although a simple encoding method can be used for text to vector conversion, the more recent methods improve the model's performance by creating token embeddings and most importantly, performing the conversion in a

context-aware manner. These token embeddings represent words with similar meanings as closely similar numeric representations while differentiating a word's representation based on the context.

The BERT model's input encoding method uses tokenized Wordpiece (Wu et al., 2016) representations of the text, segment embeddings which indicate what sentence each token belongs to, and position embeddings signifying the order of input embeddings. The input of the BERT model is an elementwise sum of these three sequences resulting in 512 parameters. Much like the input, the BERT's output is a vectored representation of the text that is referred to as BERT embeddings. The $\text{BERT}_{\text{BASE}}$ that was used in this research produces a sequence of 768 embeddings.

4.3 Crash Severity Classification Model

The main goal of this research was to develop a deep learning model to classify the severity of the crashes based on the reported narrative. Toward this goal, a transfer learning approach was undertaken using the pre-trained BERT model and its ability to create state-of-the-art models without a need for substantial task-specific architecture modifications (Devlin et al., 2019). The development and training of complex NLP models requires a substantial amount of data often in excess of billions of words, making transfer learning an effective approach. Transfer learning refers to the use of a model's previously gained knowledge to improve its capability to perform a related task. In this research, we used the Transformers open-source library (Wolf et al., 2020) to develop the Crash Severity Classification Model as depicted in Figure 1. This model has had 109,486,085 parameters, which was 3,845 more than the $\text{BERT}_{\text{BASE}}$'s 109,482,240 parameters.

After data preparation and initial preprocessing, the tokenized narrative data, segment embeddings, and position embeddings were encoded using the Transformer tokenizer's encode function (Step 2). This encoding function produced a compatible input for the pretrained BERT model. The uncased $\text{BERT}_{\text{BASE}}$ was then extended by adding a linear transformation layer corresponding to the one-hot-encoded representation of crash severity classes ("A," "B," "C," "D," and "E") (Step 3). The Model hyper-parameters were then modified based on the result obtained from train and test data subsets (Step 4). The Model was then trained, tested, and validated using the data from years 2011 to 2016 individually (Step 4). The final performance metric-based evaluation of the model was done using the test and validation subsets (Step 5).

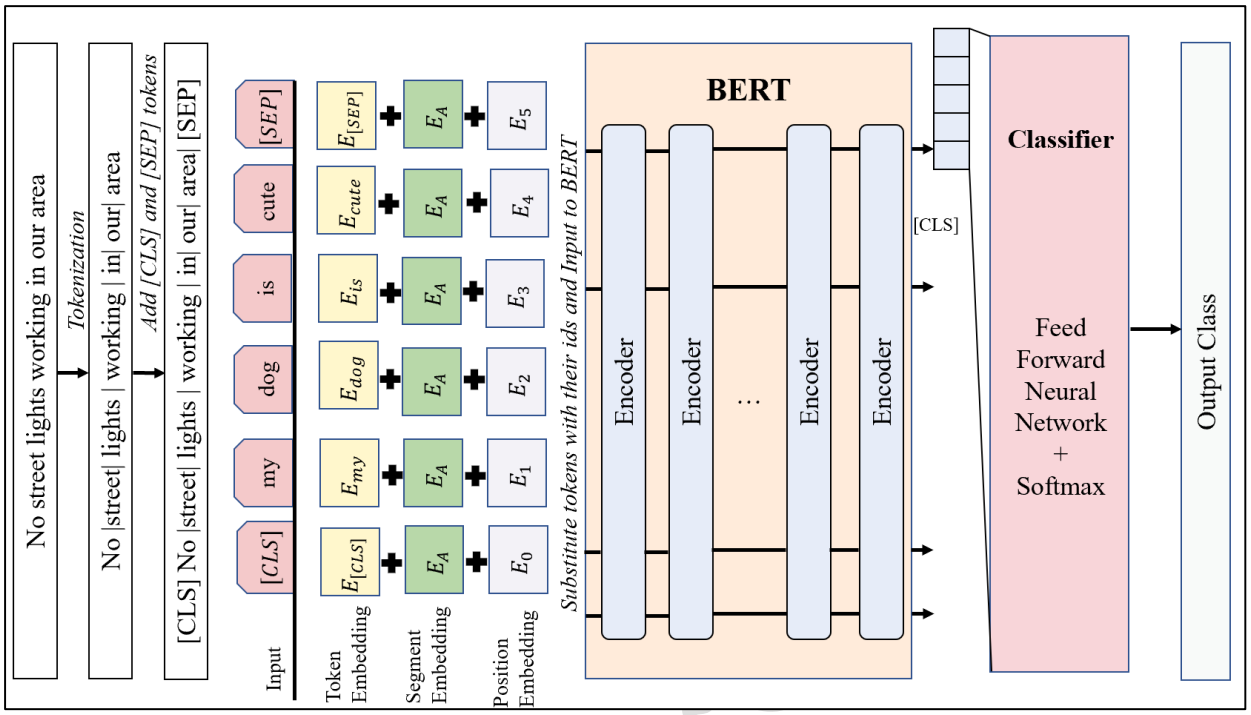


Fig. 1. Classification model network architecture.

4.4 Experimental Results and Findings

The Adam optimization algorithm (Kingma and Ba, 2022) was used in the training process to iteratively update the network weights for the classifier portion of the network. We used randomized seed values for each run of the training code.

We used the cross-entropy loss function to train the model. This function is calculated as follows:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^T, \qquad l_n = -w_{y_n} \log \frac{exp(x_{n,y_n})}{\sum_{c=1}^{C} exp(x_{n,c})}$$

Where x is the input, y is the target, w is the weight, C is the number of classes, and N spans the minibatch dimension. Here the Warmup Steps represents the percentage of overall training steps during which the learning rate increases from 0 to the Learning rate value (0.00005). The hyper-parameters used in this study are listed in Table 2.

Table 2
Hyper-parameters used in the experiments.

Hyper-parameter	Value
The dropout rate of the classification layer	0.1
Learning rate	0.00005
Epochs	6
Batch size	16
Warmup Steps	10%

4.4.1 Performance metrics

Performance metrics quantify the performance of models. The following classification metrics were used to assess the model's overall viability as a classifier and its performance (Hossin and Sulaiman, 2015):

- Accuracy (acc): The prediction accuracy of a model in the context of classification is the ratio of correct predictions over the total number of examined instances. The Accuracy is given as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (p): Precision measures positive patterns that are correctly predicted over the total positive prediction patterns. Macro Average Precision is a weighted averaged precision. These metrics are given as:

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

$$Macro\ Averaged\ Precision = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FP_i}}{l}$$

$$\text{Weighted } Averaged\ Precision = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FP_i} \times n_i}{l}$$

- Recall (r): Recall is a measure of positive patterns over the total correct predictions. Much like Precision, Recall is calculated for each class, thus, averaging is required for multiclass model assessment. This metric, its macro average, and its weighted average are given as:

$$Recall = \frac{TP}{TP + TN}$$

$$Macro\ Averaged\ Recall = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + TN_i}}{l}$$

$$\text{Weighted } Averaged\ Recall = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + TN_i} \times n_i}{l}$$

- F-measure (F1): F-measure is the harmonic mean between recall and precision values. The macro averaged F-measure and weighted Averaged recall are calculated using macro averaged and Weighted Averaged recall and precision, respectively.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Here, the samples classified by the model as positive but are negative are denoted as FP, and samples that the model classified as negative but are positive are denoted as FN. The samples that the model correctly predicted as positive and negative are denoted as TP and TN, respectively. The l represents the number of classes.

- Area under Receiver Operating Characteristics Curve (AUROC)

The area under the ROC curve is a well-known measure for classification problems. The ROC curve displays the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) of a model. The ROC is independent of classification thresholds, and the area under the ROC curve reflects the overall ranking performance of the classifier (Hossin and Subaiman, 2015; Hand. In this research, the One vs Rest (OvR) method was used to produce the ROC

curves, reducing the classification output into binary classification metrics for further evaluation.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$AUC = \frac{S_p - n_p\,(n_n + 1)/\,2}{n_p\,n_n}$$

Here, S_p is the sum of all positive samples ranked, while n_p and n_n denote the number of positive and negative samples, respectively.

5. Results and Discussions

The cross-entropy loss values for the model's training process for each year's data shows that the model could be considered fit after two or three training epochs. The main criteria for fitness assessment have been the observation of minimum cross-entropy loss for the Test subset without any further reduced loss values in the following two iterations. Early stopping of the training process prevented the model from becoming overfit. The models trained on 2011 to 2016 data subset were evaluated at their minimum Test loss (see Figure 2). The practice of fine-tuning is an iterative process that involves adjustment of hyper-parameters or modification of network structure based on the model's performance on the train and test dataset. As a result, it is possible for the researchers to produce a model that has been optimized to perform best on the test data subset be it a random sample. The validation data was thus only used after all the adjustments were applied. The observed similarity between test and validation loss (see Figure 2) confirms that metrics are representative of the model's general performance for the defined crash severity classification task.

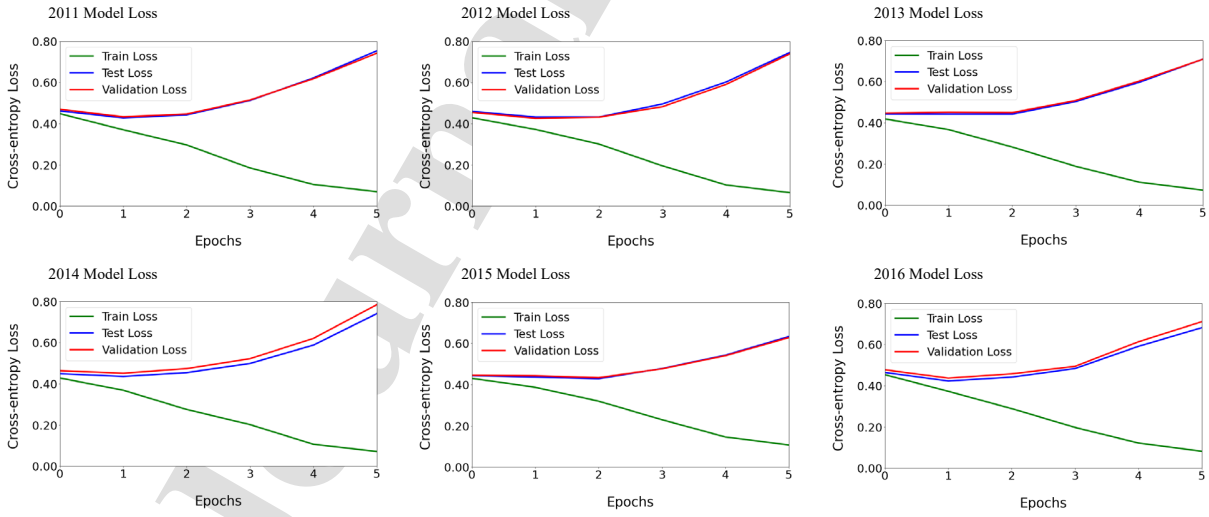


Fig. 2. Model Loss for years 2011 to 2016.

5.1. Accuracy

The assessment of classification performance metrics is a nuanced task, and a combination of metrics should be considered to discriminate between task-specific models. The model's accuracy for each year ranges from 83.67% to 84.75% for the test subset and 83.65% to 84.70% for the validation subset (Figure 3). The similarity in accuracy and loss between Test and Validation subsets suggests that results are unbiased and correctly reflect the model's classification capability. The similar proportionality in false positive and false negative predictions between Test confusion matrices (Figure 4) and their Validation counterparts (Figure 5) corroborates this finding. Besides, it can be observed that the model performs best at predicting type E crash.

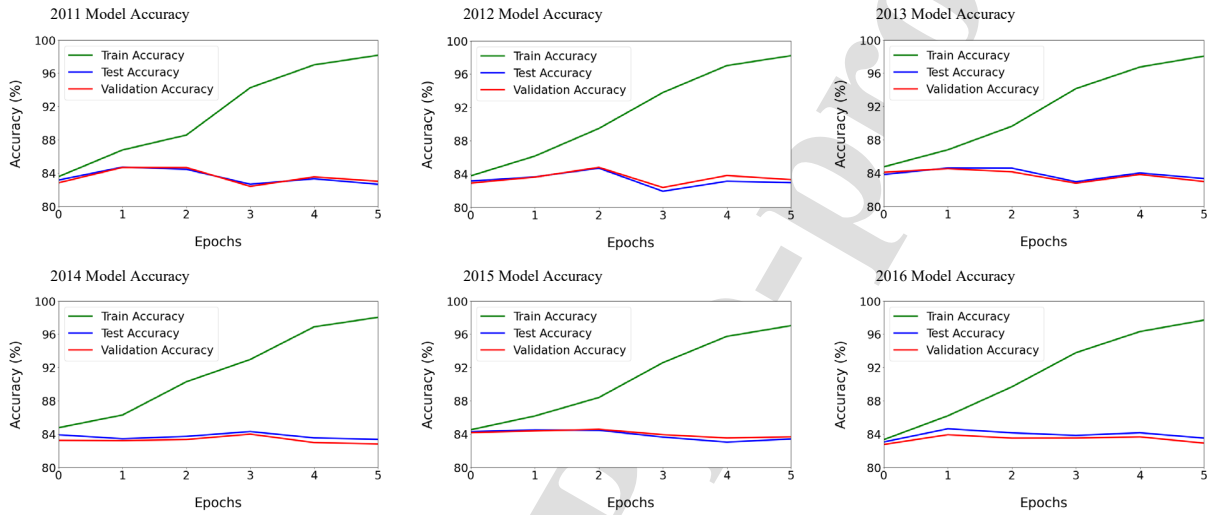


Fig. 3. Model accuracy for years 2011 to 2016.

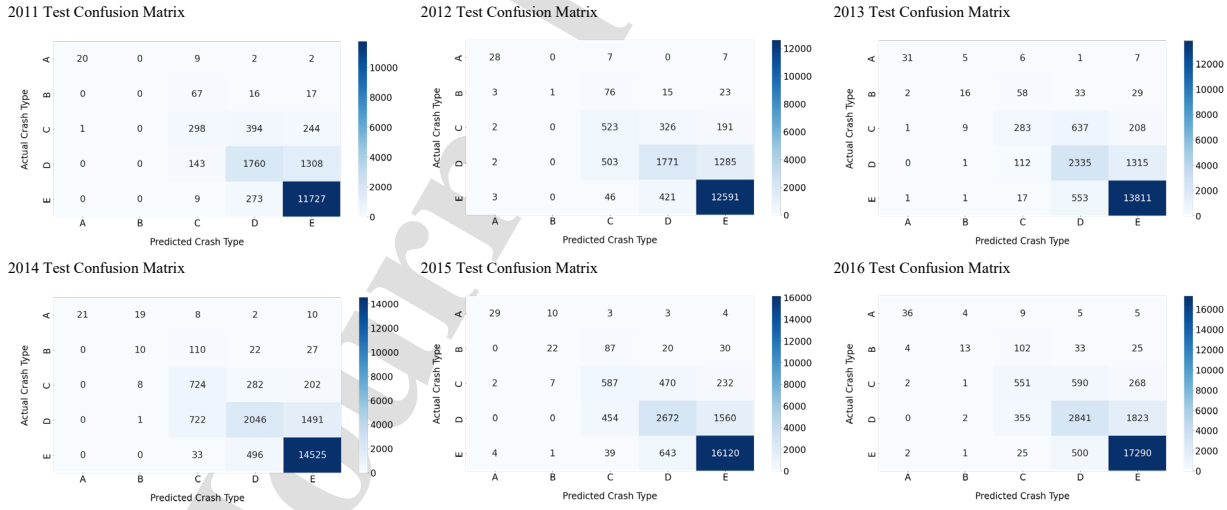


Fig. 4. Model test confusion matrix for years 2011 to 2016.

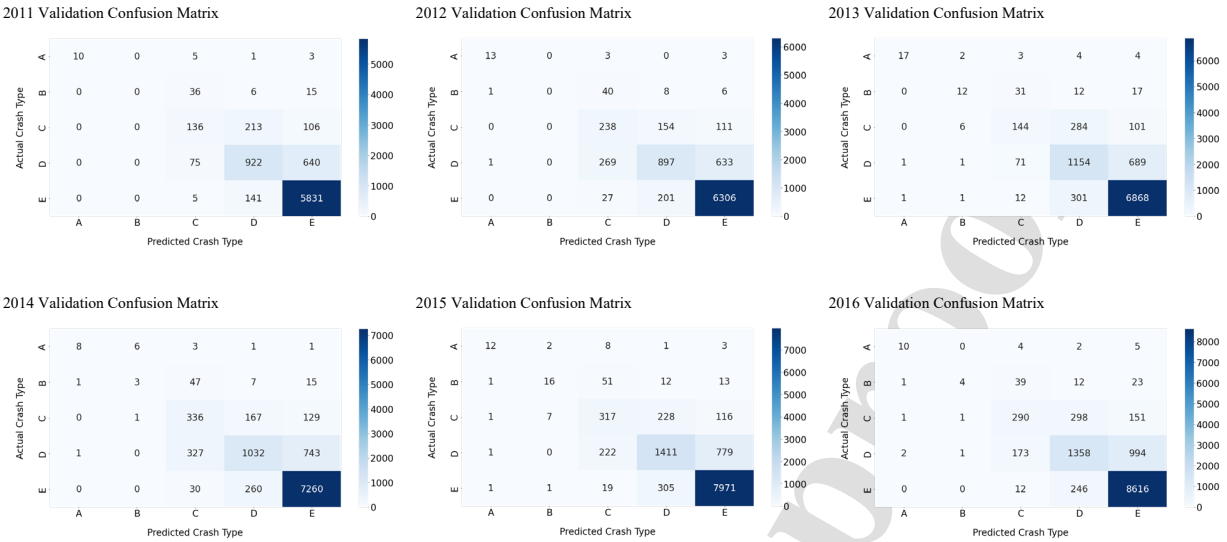


Fig. 5. Model validation confusion matrix for years 2011 to 2016.

5.2 F-measure

The weighted averaged F-measure of the model ranges from 82.52% to 83.60% for the Test subset and 82.49% to 83.73% for the Validation subset (see Table 3). This consistent performance, in contrast to the macro averaged F-measure that ranges from 53.94% to 60.37%, points to the fact that the model's classification performance is not consistent for each class. For instance, the "B" crash severity class is at times completely neglected (2011 and 2012 Test and Validation confusion matrix: Figure 5, Figure 6) or mostly misclassified (2013 to 2016 Test and Validation confusion matrix). The model's capability to correctly identify each class does not simply correspond to the proportional representation of the class in the training dataset. It also corresponds to the distinctiveness of the class's sentiment of narratives. For instance, the model can identify the class "A" narratives more accurately in comparison to that of class "B." Model accuracies for individual crash type is listed in Table 4.

Table 3
Model performance for individual years.

Dataset	Subset	Support	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
2011	Test	16290	84.7%	62.4%	49.0%	53.9%	82.7%	84.8%	83.1%
	Valid[1]	8145	84.7%	62.6%	47.3%	52.6%	82.5%	84.7%	83.1%
2012	Test	17824	83.7%	75.6%	52.8%	54.0%	82.9%	83.7%	82.5%
	Valid	8911	83.7%	57.7%	52.4%	54.4%	82.4%	83.7%	82.5%
2013	Test	19472	84.6%	70.7%	51.3%	56.7%	83.1%	84.6%	83.3%
	Valid	9736	84.2%	70.9%	51.2%	57.3%	82.6%	84.2%	83.0%
2014	Test	20759	84.3%	70.3%	56.4%	60.4%	83.0%	84.3%	83.2%
	Valid	10378	84.0%	65.4%	56.6%	58.8%	82.3%	84.0%	82.8%
2015	Test	22999	84.5%	69.6%	54.2%	58.9%	83.3%	84.5%	83.6%
	Valid	11498	84.6%	70.0%	53.1%	58.1%	83.6%	84.6%	83.7%
2016	Test	24487	84.7%	71.5%	52.2%	56.8%	83.2%	84.7%	83.4%
	Valid	12243	84.0%	70.6%	48.6%	53.2%	82.4%	84.0%	82.5%

Note: [1]Valid= Validation

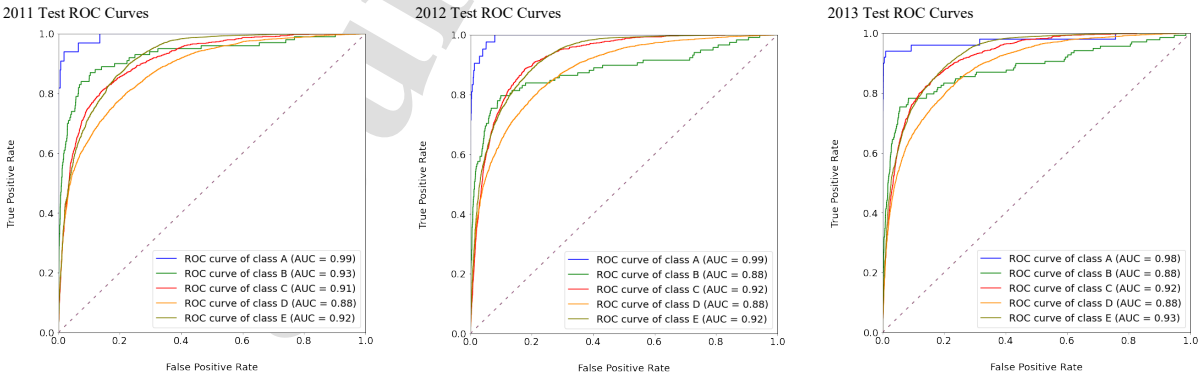
Table 4

Model accuracy for individual crash severity type.

Dataset	Subset	A	B	C	D	E
2011	Test	99.91%	99.39%	94.68%	86.89%	88.62%
	Validation	99.89%	99.30%	94.60%	86.79%	88.83%
2012	Test	99.87%	99.34%	93.54%	85.68%	88.91%
	Validation	99.91%	99.38%	93.22%	85.79%	88.99%
2013	Test	99.88%	99.29%	94.62%	86.38%	89.06%
	Validation	99.85%	99.28%	94.78%	86.00%	88.43%
2014	Test	99.81%	99.10%	93.42%	85.47%	89.12%
	Validation	99.87%	99.26%	93.22%	85.49%	88.65%
2015	Test	99.89%	99.33%	94.37%	86.30%	89.07%
	Validation	99.84%	99.24%	94.33%	86.54%	89.24%
2016	Test	99.87%	99.30%	94.48%	86.49%	89.18%
	Validation	99.88%	99.37%	94.45%	85.89%	88.31%

5.3 AUROC

As it is often the case that crash severity datasets are typically imbalanced, there is usually a trade-off between evaluation measures that based on one specific threshold. In contrast to the previous metrics which evaluate the model based on the model's class predictions, the ROC is drafted and AUROC is calculated based on the model's predicted probability of class membership. Its area under the curve (AUC) represents the degree of separability between classes. With a maximum of ROC-AUC value close to 1 describing that the classifier has an excellent performance in separating classes, and a value close to 0.5 describing a valueless test. The model's OvR AUROC performance for the Test subset ranges from 0.88 to 0.99 (Figure 6), a range similar to that of the Validation subset (Figure 7). This points to the fact that the model in general assigns high probability for the correct class, even though the highest assigned probability might be incorrect more often for certain classes.



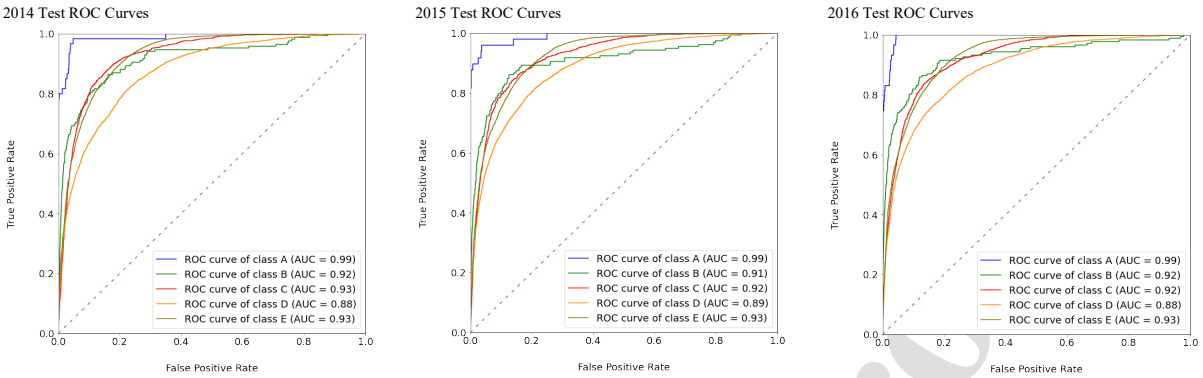


Fig. 6. Model test ROC curves for years 2011 to 2016.

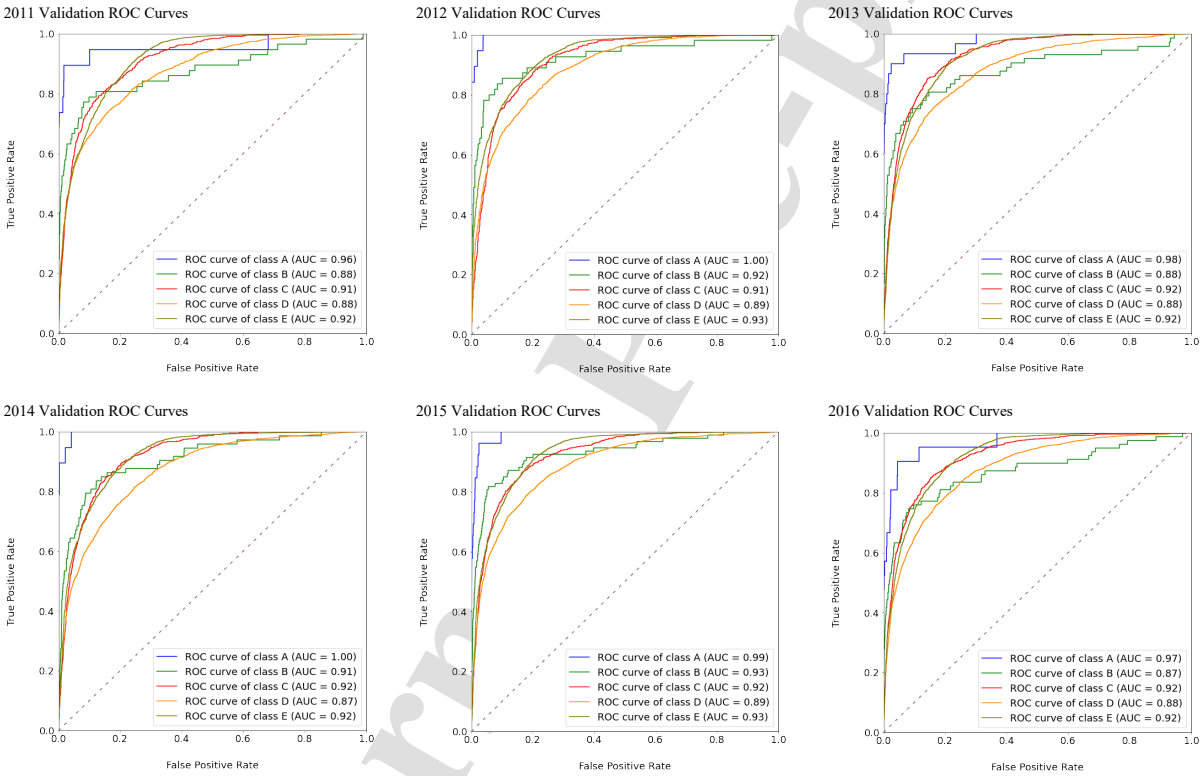


Fig. 7. Model validation ROC curves for years 2011 to 2016.

6. Conclusions

This study uses a large sample of crash narrative data to apply BERT advanced language model to classify crash severity types accurately. The research presents an advanced method for improving prediction accuracy from noisy crash narrative data. More importantly, the study contributed to developing a framework that could be used to address other transportation safety-related classification issues in the future with high accuracy. This study's results can help predict the severity of a reported crash with unknown severity and identify additional factors that contribute to crashes.

The proposed model's architecture consists of a pretrained $\text{BERT}_{\text{BASE}}$ embedding network and a linear transformation layer that we finetuned. The model was developed using six years (2011-

2016) of traffic crash data from Louisiana. The elevated level of noise in the narrative data necessitated several preprocessing measures. The model was trained using 70% of the data, 20% to adjust the hyper-parameters, and the remaining 10% to validate the findings. The model's overall Accuracy was above 83.65%, and the weighted averaged F-measure was above 82.52%. The Macro averaged F-measure of the model was above 53.94%, suggesting that the model's performance was lower for the minority classes and the classes with harder-to-distinguish sentiments. These findings encourage the exploration of other text encoding methodologies, data augmentation approaches, hyper-parameter adjustment, and fundamental changes in neural network architecture as improvement tactics. Beyond model-related advances, forming more balanced datasets with lower noise could help assess future models and increase their real-world applicability.

The current study is not without limitations. The data used in this study included five crash severity classes of differing sizes. The lower accuracy for minority classes limited the model's generalization capability. For future research, we recommend the incorporation of under-sampling and data augmentation techniques as possible solutions for this issue. The proposed network architecture could be applied to other highway safety tasks, including text-based classification steps. Additionally, the dataset is not recent. The research team was limited to using textual content available to them for usage. Future studies can collect recent crash narrative data to explore the performance of the recently documented crash narrative reports.

CRediT authorship contribution statement

The authors confirm contribution to the paper as follows: study conception and design: S. Das; data collection: S. Das; analysis and interpretation of results: S. Das, A. Oliaee; draft manuscript preparation: S. Das, A. Oliaee, J. Liu, and M. A. Rahman; All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Bareiss, M., C. Smith, and H. C. Gabler. Finding And Understanding Pedal Misapplication Crashes Using A Deep Learning Natural Language Model. *Traffic Injury Prevention*, Vol. 22, No. S1, 2021, p. pp S169-S172. https://doi.org/10.1080/15389588.2021.1982616.

Bunn, T. L., S. Slavova, and L. Hall. Narrative Text Analysis Of Kentucky Tractor Fatality Reports. *Accident Analysis & Prevention*, Vol. 40, No. 2, 2008, p. pp 419-425.

Chawla, H., M.-U. Megat-Johari, P. T. Savolainen, and C. M. Day. Evaluation of Strategies to Mitigate Culvert-Involved Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2675, No. 5, 2021, p. pp 403-417. https://doi.org/10.1177/0361198121992070.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv Preprint ArXiv:2003.10555*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, 2978–2988. https://doi.org/10.18653/v1/P19-1285

Das, S. Understanding Fatal Crash Reporting Patterns in Bangladeshi Online Media Using Text Mining. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2675, No. 10, 2021, p. pp 960-971. https://doi.org/10.1177/03611981211014200.

Das, S., A. Dutta, and I. Tsapakis. Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2675, No. 9, 2021, p. pp 449-462. https://doi.org/10.1177/03611981211002523.

Das, S., M. Le, and B. Dai. Application of Machine Learning Tools in Classifying Pedestrian Crash Types: A Case Study. *Transportation Safety and Environment*, Vol. 2, No. 2, 2020, p. pp 106-119. https://doi.org/10.1093/tse/tdaa010.

Das, S., S. Datta, H. A. Zubaidi, and I. A. Obaid. Applying Interpretable Machine Learning To Classify Tree And Utility Pole Related Crash Injury Types. *IATSS Research*, Vol. 45, No. 3, 2021, p. pp 310-316. https://doi.org/10.1016/j.iatssr.2021.01.001.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, 2019.

Fitzpatrick, C. D., S. Rakasi, and M. A. Knodler. An Investigation Of The Speeding-Related Crash Designation Through Crash Narrative Reviews Sampled Via Logistic Regression. *Accident Analysis & Prevention*, Vol. 98, 2017, p. pp 57-63. https://doi.org/10.1016/j.aap.2016.09.017.

Gao, L., H. Wu, and Transportation Research Board. Verb-Based Text Mining of Road Crash Report. 2013.

Goh, Y. M., and C. U. Ubeynarayana. Construction Accident Narrative Classification: An Evaluation Of Text Mining Techniques. *Accident Analysis & Prevention*, Vol. 108, 2017, p. pp 122-130. https://doi.org/10.1016/j.aap.2017.08.026.

Hand, D. J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. p. 16.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv Preprint ArXiv:2006.03654*.

Hossin, M., and SulaimanM.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, Vol. 5, No. 2, 2015, pp. 01–11. https://doi.org/10.5121/ijdkp.2015.5201.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention 108, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jiang, Z.-H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). Convbert: Improving bert with span-based dynamic convolution. Advances in Neural Information Processing Systems, 33, 12837–12848.

COMPRESSING TEXT CLASSIFICATION MODELS. 2017, p. 13.

Kim, J., A. B. Trueblood, H.-C. Kum, and E. M. Shipp. Crash Narrative Classification: Identifying Agricultural Crashes Using Machine Learning with Curated Keywords. *Traffic Injury Prevention*, Vol. 22, No. 1, 2021, p. pp 74-78. https://doi.org/10.1080/15389588.2020.1836365.

Kingma, D. P., and J. Ba. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980. Accessed Jun. 9, 2022.

Krause, S., & Busch, F. New insights into road accident analysis through the use of text mining methods. (2019, June). *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (pp. 1-6). IEEE.

Kwayu, K. M., V. Kwigizile, J. Zhang, and J.-S. Oh. Semantic N-Gram Feature Analysis and Machine Learning–Based Classification of Drivers' Hazardous Actions at Signal-Controlled Intersections. *Journal of Computing in Civil Engineering*, Vol. 34, No. 4, 2020, p. Content ID 04020015. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000895.

Kwayu, K. M., V. Kwigizile, K. Lee, and J.-S. Oh. Discovering Latent Themes In Traffic Fatal Crash Narratives Using Text Mining Analytics And Network Topology. *Accident Analysis & Prevention*, Vol. 150, 2021. https://doi.org/10.1016/j.aap.2020.105899.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. http://arxiv.org/abs/1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Lopez, D., L. C. Malloy, and K. Arcoleo. Police Narrative Reports: Do They Provide End-Users With The Data They Need To Help Prevent Bicycle Crashes? *Accident Analysis & Prevention*, Vol. 164, 2022.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. p. 9.

Møller, M., K. H. Janstrup, and N. Pilegaard. Improving Knowledge Of Cyclist Crashes Based On Hospital Data Including Crash Descriptions From Open Text Fields. *Journal of Safety Research*, Vol. 76, 2021, p. pp 36-43. https://doi.org/10.1016/j.jsr.2020.11.004.

Pennington, J., R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. Presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Sayed, M. A., X. Qin, R. J. Kate, D. M. Anisuzzaman, and Z. Yu. Identification And Analysis Of Misclassified Work-Zone Crashes Using Text Mining Techniques. *Accident Analysis & Prevention*, Vol. 159, 2021. https://doi.org/10.1016/j.aap.2021.106211.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). *MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices* (arXiv:2004.02984). arXiv. http://arxiv.org/abs/2004.02984

Taylor, W. L. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, Vol. 30, No. 4, 1953, pp. 415–433. https://doi.org/10.1177/107769905303000401.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wali, B., A. J. Khattak, and N. Ahmad. Injury Severity Analysis Of Pedestrian And Bicyclist Trespassing Crashes At Non-Crossings: A Hybrid Predictive Text Analytics And Heterogeneity-Based Statistical Modeling Approach. *Accident Analysis & Prevention*, Vol. 150, 2021. https://doi.org/10.1016/j.aap.2020.105835

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz. Transformers: State-of-the-Art Natural Language Processing. 2020

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* (arXiv:1609.08144). arXiv. http://arxiv.org/abs/1609.08144

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (arXiv:1906.08237). arXiv. http://arxiv.org/abs/1906.08237

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., & Yang, L. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297.

Zhang, X., P. Srinivasan, and S. Mahadevan. Sequential Deep Learning From NTSB Reports For Aviation Safety Prognosis. *Safety Science*, Vol. 142, 2021, p. 105390. https://doi.org/10.1016/j.ssci.2021.105390.

Zhang, X., Green, E., Chen, M., & Souleyrette, R. R.. Identifying secondary crashes using text mining techniques. (2020). *Journal of Transportation Safety & Security*, *12*(10), 1338-1358.

Using Bidirectional Encoder Representations from Transformers (BERT) to Classify Traffic Crash Severity Types

Amir Hossein Oliaee, Ph.D. Candidate
College of Architecture, Texas A&M University
Langford Architecture Building 3137, College Station, TX 77840
Email: oliaee@tamu.edu
ORCID: 0000-0002-0574-2690

Subasish Das, Ph.D.
(Corresponding author)
Assistant Professor, Ingram School of Engineering
Texas State University
601 University Drive, San Marcos, Texas 78666
E-mail: subasish@txstate.edu
ORCID ID: 0000-0002-1671-2753

Jinli Liu, Ph.D. Student
College of Liberal Arts, Texas State University
601 University Drive, San Marcos, Texas 78666
Email: j_l848@txstate.edu
ORCID: 0000-0002-6152-8808

M. Ashifur Rahman, Ph.D.
Research Associate, University of Louisiana at Lafayette
E-mail: ashifur@louisiana.edu
ORCID: 0000-0001-6940-1599

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.