Using Text Data from the DT4000 to Enhance Crash Analysis

Xiao Qin, Ph.D., PE
Rohit Kate, Ph.D.
Md Abu Sayed, MS
D M Anisuzzaman, MS

Institute for Physical Infrastructure and Transportation (IPIT)
University of Wisconsin-Milwaukee

WisDOT ID FG-2020-UW-MILWA-05069 and FG-2021-UW-MILWA-05641

September 2021

WISCONSIN DEPARTMENT OF TRANSPORTATION

RESEARCH & LIBRARY UNIT

WISCONSIN DOT
PUTTING RESEARCH TO WORK

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Using Text Data from the DT4000 to Enhance Crash Analysis		**5. Report Date** September 2021
		6. Performing Organization Code
7. Author(s) Xiao Qin, Rohit Kate, Md Abu Sayed, D M Anisuzzaman		**8. Performing Organization Report No.**
9. Performing Organization Name and Address Institute for Physical Infrastructure and Transportation (IPIT) University of Wisconsin-Milwaukee Milwaukee, WI 53201-0784		**10. Work Unit No.**
		11. Contract or Grant No. FG-2020-UW-MILWA-05069 and FG-2021-UW-MILWA-05641
12. Sponsoring Agency Name and Address Wisconsin Department of Transportation Research & Library Unit 4822 Madison Yards Way Room 911 Madison, WI 53705		**13. Type of Report and Period Covered** Final Report October 2019-September 2021
		14. Sponsoring Agency Code

15. Supplementary Notes

If applicable, enter information not included elsewhere, such as translation of (or by), report supersedes, old edition number, alternate title (e.g. project name), or hypertext links to documents or related information.

16. Abstract

Eighty percent of the world's data is unstructured, meaning it lives in text documentation, photos, audio files and videos files. Unstructured data cannot be easily stored in a database, and even if it is stored, it has attributes that make it a difficult to edit, query and analyze, especially on the fly. Crash narratives fall into this category, meaning they are not adequately captured in many of the current analyses that are being used to guide highway planning and design and make driving safer. For example, law enforcement officers' narratives can contain valuable insight, but because it is not necessarily included in structured data fields and attributes, it's challenging information for traffic safety engineers to query and obtain. Additionally, keyword searches often return irrelevant results (false positives) because words often have multiple meanings. On the other hand, a language has rich vocabularies that allow people to say things differently. While engineers often manually review these reports for causes and contributing factors that might guide remedial actions, the process is labor intensive, and the review quality is inconsistent as it is subject to the reviewers' experience and judgement. This study developed, tested and implemented intelligence algorithms (i.e., natural language processing, text mining, statistical modeling) that extract data from police narratives.

Six state-of-the practice machine learning (ML) based classifiers were applied to crash narratives collected in Wisconsin from 2017 to 2021: multinomial naive bayes (MNB), logistic regression (LGR), support vector machine (SVM), random forest (RF), K-nearest neighbor (K-NN), recurrent neural network with Gated Recurrent Unit (GRU); and one probabilistic classifier - NoisyOR. Creating the training data for these methods did not require much manual annotation work because much of it could be automatically created by leveraging the structured portion of the crash reports. The classifiers were evaluated experimentally for two case studies: 1) identifying missed work zone crashes; and 2) identifying missed crashes related to distracted driving and inattentive driving and separating the two if a case is reported as both. The evaluation of the seven classifiers consistently identified NoisyOR as the top classifier. The strengths and limitations of applying text mining techniques to analyze crash narratives were further discussed, leading to insights into the nature of crash narratives and how highway safety analysis can be best benefited by using these techniques.

17. Key Words Crash Data, DT4000, Crash Narrative, Crash Analysis, Text Mining, Machine Learning, Unstructured Data	**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 61	**22. Price**

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

This research was funded by the Wisconsin Department of Transportation through the National Highway Traffic Safety Administration (NHTSA) under project FG-2020-UW-MILWA-05069 & FG-2021-UW-MILWA-05641. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views of the Wisconsin Department of Transportation (WisDOT) or National Highway Traffic Safety Administration (NHTSA) at the time of publication.

This document is disseminated under the sponsorship of the Wisconsin Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. This report does not constitute a standard, specification or regulation.

The United States Government does not endorse products or manufacturers. Trade and manufacturers’ names appear in this report only because they are considered essential to the object of the document.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Motor vehicle accident reports are the most useful and valuable source for analyzing crashes and identifying factors that contribute to crashes. Most information from a crash is categorized as “structured” data in that it is provided by law enforcement agencies through crash report forms with appropriate data fields that are then stored in a database. However, law enforcement officers also provide a significant amount of detailed information in the crash narrative, which is presented in an unstructured text format. The narrative fields can be used to record additional information on specific circumstances, key factors (e.g., citations, additional witnesses, types of drugs and medication, hazardous materials spillage from trucks and buses, trailer and towed, school bus information) and other circumstances that could be leading to a crash but is not compiled through structured data fields. More importantly, the narrative provides detailed explanations to these contributing circumstances such as driver, vehicle or highway; and often times, specific crash location information. Unstructured data can’t be easily stored in a database. And even if it is stored, it has attributes that make it a difficult to edit, query and analyze, especially on the fly.

However, structured data from crash reports doesn’t provide perfect information either. For example, incorrectly classifying a crash will lead to undercounting some types of crashes and overcounting others. Crashes might also be completely missing from structured data fields due to: restrictive reporting options in tabular forms (Blackman, Debnath, and Haworth 2020; Ullman and Scriba 2004; Wang et al. 1996); lack of understanding about the importance of the crashes, overloaded by work during crash reporting time (Graham and Migletz 1983); and misclassification of crashes (Wang et al., 1996, Farmer, 2003). Furthermore, generally, a police officer makes certain judgments about a crash based on the severity of the crash and based on the driver. For example, a fatal crash is usually given the highest reporting priority, compared with property damage crashes, which usually receive a lower priority (Ye and Lord 2011). Furthermore, crashes with less severity or with no injuries are sometimes not even reported in structured data (Wang et al. 1996). Additionally, the probability of reporting an injury crash increases with the number of vehicles involved as well as the age of the injured (i.e., crashes with young children are reported 20-30% of the time, and crashes with persons over 60 are reported 70% of the time) (Hauer and Hakkert 1988). Also, crashes involving younger or female drivers have a lower probability of being reported (Amoros, Martin, and Laumon 2006). Therefore, estimates based solely on structured data fields do not provide complete information for the safety analysis, meaning crash narratives are an important piece of the puzzle (Abay 2015).

Manual review of crash narratives for causes and contributing factors, while immensely valuable, is labor-intensive for traffic safety engineers because the language varies by report. The results are also somewhat inconsistent because they are subject to the reviewers’ experience and judgement. Therefore, a more predictable, consistent, and efficient method of automatic information extraction, such as text mining, is necessary. A crash narrative can be converted to a numeric vector suitable for machine learning, a process often referred to as feature extraction.

Text mining results can be used to assess the quality of crash flags (e.g., work zone, secondary crash) and identify vehicle actions in a crash by the sequence of verb phases in a narrative. Moreover, thousands of crash reports can be reviewed in a matter of minutes using these techniques, according to a recent study.

The goal of this study is to develop intelligent algorithms (i.e., machine learning (ML) natural language processing (NPL), text mining, statistical modeling) that facilitate the rapid and efficient retrieval of critical data from police crash narratives.

This goal will be achieved through the following objectives:

1) Perform a comprehensive literature review with the emphasis on machine learning and text mining techniques and their applications. Particular attention will be given to applications that have been developed for analyzing traffic accident reports.
2) Conduct interviews with traffic engineers and safety practitioners who have experience reviewing and analyzing crash reports. The interviews can help identify the challenges of obtaining quality information from narratives, understand which information in the narratives is more frequently searched and identify the methods utilized by analysts to process information from crash narratives.
3) Collect 3-5 years of crash data, including crash narratives, from WisTransportal and identify appropriate case studies to demonstrate how the application of text mining advances crash analyses.
4) Develop machine learning and text mining algorithms using different methodologies.
5) Compare the algorithms' performance based on two case studies (i.e., crashes related to work zones and crashes related to distracted or inattentive driving) and make recommendations for improving crash data quality and enhancing safety analysis.

This report documents the text mining techniques and tools, assessment outcomes of crash flags (e.g., work zone), and the key contextual information relating to a crash. Such information can be vital for seeking the causal factors of a crash. In addition, maximizing the value of a crash narrative will encourage and incentivize law enforcement agencies to use a narrative to capture data that are not available in the data fields.

2. LITERATURE REVIEW

Text mining was introduced as a way to enable machine-supported analysis of text (Feldman and Dagan 1995). Information retrieval, natural language processing, information extraction, text summarization, opinion mining and sentiment analysis are some of important areas of text mining research (Allahyari et al. 2017). Text mining has become both popular and necessary in many fields, including financial services, health care, transportation, communication and media, information technology and internet, political analysis, public administration and legal services (Gupta, Lehal, and others 2009; Inzalkar and Sharma 2015; Maheswari and Sathiaseelan 2017).

Natural Language Processing (NLP) is a computerized text mining technology used for analyzing text data, which is based on a set of theories and technologies for the purpose of achieving human like language processing (Liddy 2001). An important branch of NLP is text classification which aims to assign labels to narratives based on the content or context of the narratives. Recently, text classification attained considerable attention because of its application in e-mail filtering, spam detection, web-page content filtering, automatic message routing, automated indexing of articles, and searching for relevant information on the Web (Kowsari et al. 2019). Meanwhile, a variety of machine learning (ML) techniques have been developed, such as Bayesian classifiers, support vector machines, k nearest neighbors, decision trees, and neural networks and have been successfully used for text classification for decades (Yang 1999). The growing interest in NLP and ML techniques, as well as the availability of textual dataset has prompted the applications of such techniques for text classification in transportation engineering fields, especially in the highway safety data analysis. Some of the notable safety applications include (1) crash contributing factor identification, (2) crash severity analysis, (3) crash event and cause analysis, and (4) crash type classification (e.g., speed-related, pedestrian-related). The dataset used to classify crash narratives includes various text data such as police reports, auto insurance claims reports, and social media data.

Most text mining algorithms require some text preprocessing, such as tokenization, filtering, lemmatization, stemming, etc. Once preprocessing has been completed, algorithms for classification, clustering, or information extraction are applied to the text. Some commonly used clustering algorithms are hierarchical clustering, k-means clustering, and probabilistic clustering and topic models (e.g., probabilistic latent semantic analysis, latent Dirichlet allocation) (Allahyari et al. 2017). Examples of popular classification algorithms include naive Bayes, nearest neighbor, decision tree, decision rule, support vector machine, logistic regression, Rocchio’s algorithm, neural network, associative classifier, and centroid based classifier (Allahyari et al. 2017; Brindha, Prabha, and Sukumaran 2016; Korde and Mahender 2012).

In highway safety analysis, most of the text mining-based studies are conducted using social media and medical data, while a few studies are conducted using crash narratives. Text mining techniques used to identify a specific type of crashes are primarily based on keywords, or words that are direct or indirect indicators of certain unique and specific crash characteristics. For example, Sorock et al. applied Haddon’s injury epidemiology model of crash phases to identify pre-crash vehicle activities and various work zone crashes from automobile insurance claim narratives. In a pilot study, they manually selected a set of work zone-related words and showed that the keyword “construction” had maximum frequency in the dataset (Sorock, Ranney, and Lehto 1996). Williamson et al. extracted patterns of events of fatal injuries from crash narratives based on a pre-established text search mechanism (Williamson et al. 2001).

Many researchers used keywords-based text analysis in transportation safety. For example, Zheng et al. identified secondary crashes by using the keywords, and the distance of keywords, which was calculated by the absolute difference of indexes between two types of keywords: relationships keywords (RKWs) and events keywords (EKWs) (Zheng et al. 2015). Rakotonirainy et al. used a keyword selection approach that automatically selects keywords in the narratives. They used text mining to identify curve-related crash factors and their associated severity from insurance claim reports. The words mentioned only in curve-related crashes were selected as keywords, and the keywords with high frequencies were used as the main factors contributing to curve-related crashes (Rakotonirainy et al. 2015). Gao and Wu developed a verb-based text mining method by applying various Natural Language Processing (NLP) techniques that automatically identify the sequence of crash events from crash narratives (Gao and Wu 2013). Their method utilized syntactic and semantic information from the text to overcome the limitations of their previous methods that used predefined keywords. However, the process was not completely automatic, as the words with similar meaning had to be grouped together manually. Trueblood et al. developed a classifier tool in Excel to identify agricultural crash from crash narrative. The authors prepared two lists of keywords (agricultural and nonagricultural) manually and used the lists to search keywords in the narratives for identifying the agricultural crashes (Trueblood et al. 2019). However, their classifier assigns equal weight to the narratives that are related to agricultural crash, so it may not be effective for large data sets in which narratives are more relevant to agricultural crash.

With the advance in computational technology, the use of machine learning techniques for text mining is also noticeable. Nayak et al. applied Bayesian theory Leximancer tool to find the major contributing factors of crashes from crash narratives (Nayak, Piyatrapoomi, and Weligamage 2009). Zhang et al. conducted a comparative study on Naive Bayes, SVM and decision tree methods to find hazardous behaviors from the crash narratives reported by police, and found that Naive Bayes is the best binary classifier (Zhang, Kwigizile, and Oh 2016). However, the process is not fully automatic because samples of the crash narrative are randomly selected, and the samples are manually annotated for training and testing the model. Williams et al. applied latent semantic analysis (LSA) and latent dirichlet allocation (LDA) text mining techniques, a bayesian model, to detect accident from narratives (Williams and Betak 2018). McAdams et al. used multivariate logistic regression to study the role of helmets in reducing the injury rate of bi-vehicles using narratives collected from the national electronic injury Surveillance (McAdams et al. 2018). The narrative describes the events of actions occurred at the time of accident, and these events help determine the importance of helmet use. Heidarysafa et al. used various combinations of deep learning models to study the use of text narratives in finding the causes of accidents (Heidarysafa et al. 2019). In order to check whether the reported crash causes are consistent with the narrative description, the authors used various combinations of one-dimensional convolutional neural netwroks (CNN) and recurrent neural networks (RNN) (with LSTM and GRU units) with word2vec and GLoVe word embeddings (add citation). Although CNN and RNN with word2vec provided better results compared to other base models, these models did not work well for minority classes in the dataset. Fitzpatrick et al. used logistic regression to identify speed-related missed crash from noisy crash reports (Fitzpatrick, Rakasi, and Knodler 2017). But their data processing technique is not completely automatic and requires other secondary data (such as structure and road inventory data) to eliminate data noise. Das, S. et al. used SVM, RF and XGBoost technologies to classify pedestrian crashes from text narratives and found that XGBoost has an accuracy rate of 72%, which is more accurate than the other two models (Das, Le, and Dai 2020). They used a very small dataset to train and test the model, which may weaken the validity of the results.

The above discussion shows that naive Bayes, logistic regression, decision tree, SVM, and RNN are some of the commonly used classifiers in the highway safety field, but their limitations affect the performance of the model in varying degrees. Similar to finding speed-related or pedestrian crashes, the problem of identifying missed work-zone (WZ), distracted (DD) and inattentive (ID) crashes can be solved by assigning WZ, DD and ID labels to relevant crash narratives, which is essentially a text classification task and hence can be automated. Although there are other textual data sources such as insurance claims, social media, and news reports, the crash reported by the police is more useful because it can be used in conjunction with structural and graphical data. In addition, police officers follow general guidelines to record crashes despite that the narrative structure varies by officer, location, time, environmental conditions, and crash severity.

While past research has focused on analyzing various aspects of traffic crashes from crash narratives, none of the studies emphasized missed crashes. Their methods are either complicated, time-consuming, external data dependent or require substantial manual intervention, which does not meet our research goals. Moreover, the data from other sources are not easily accessible to the public and can be costly. In this project we implemented and thoroughly evaluated multiple text classification methods, including a new classifier Noisy-OR, and compared them. We used text classification to automate identification of missed crashes from crash narratives. This task is particularly challenging due to data quality issues.

3. WISDOT DT4000 CRASH REPORT NARRATIVE SURVEY

By targeting both groups of the collectors and users, a comprehensive survey was created to pinpoint what works and does not. Working with WisDOT, an initial set of questions was defined to be distributed. The survey contained 12 questions for the data user and 7 questions for the data collector. The survey results in this report are based on 16 responses of Data Users and Data Users / Collectors. Most questions were answered on every survey taken. If the responder selected that they are a data user & data collector, they answered both sets of questions.

Of the 16 who responded, 6 (37.5%) were identified as a government employee/onsite consultant and 7 (43.75%) are consultants funded by the government. The individual who answered other identified as a consultant that uses the information for projects funded by both the government and other agencies.

81% responders said when reviewing the crash data report, 75-100% times they need to review the crash narrative section, which shows how important a crash narrative to safety analysis. The following **Error! Reference source not found.** shows the type of work the responders need to r eview the crash narratives. As can be seen, the top three reasons are highway safety improvement program, safety review for improvement programs; and citizen requests. When other was indicated, specific reference was made to planning or studies in three of the four other responses. One response indicated the data was used for safety certification documentation for projects that include safety flags.

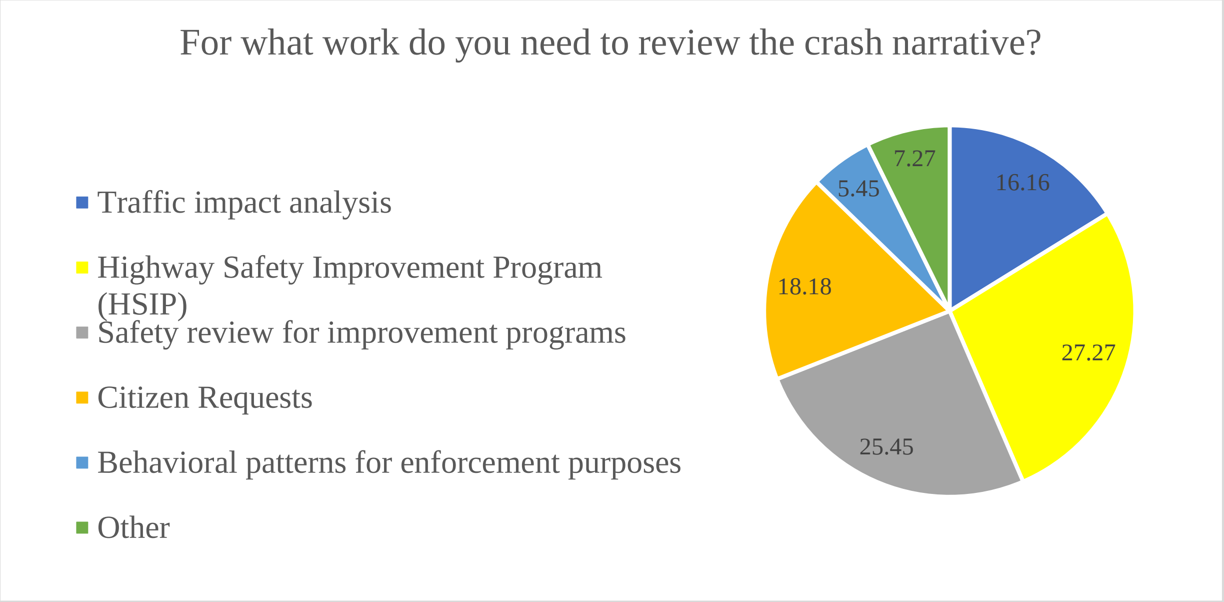

Figure 3-1 Survey Result: Use of Crash Narratives

When asked “Do you look for specific and detailed location information such as which side of the roadway or which approach? If you do, please rate the sufficiency of the location information saved in the data fields on a 1-5 scale with 5 being the best.”. Of the 16 respondents, all 16 indicated that they look for specific information within the narrative. The average of the responses is 3.38.

The following **Error! Reference source not found.** shows the reasons why they need to review t he crash narrative section given the structured data fields.

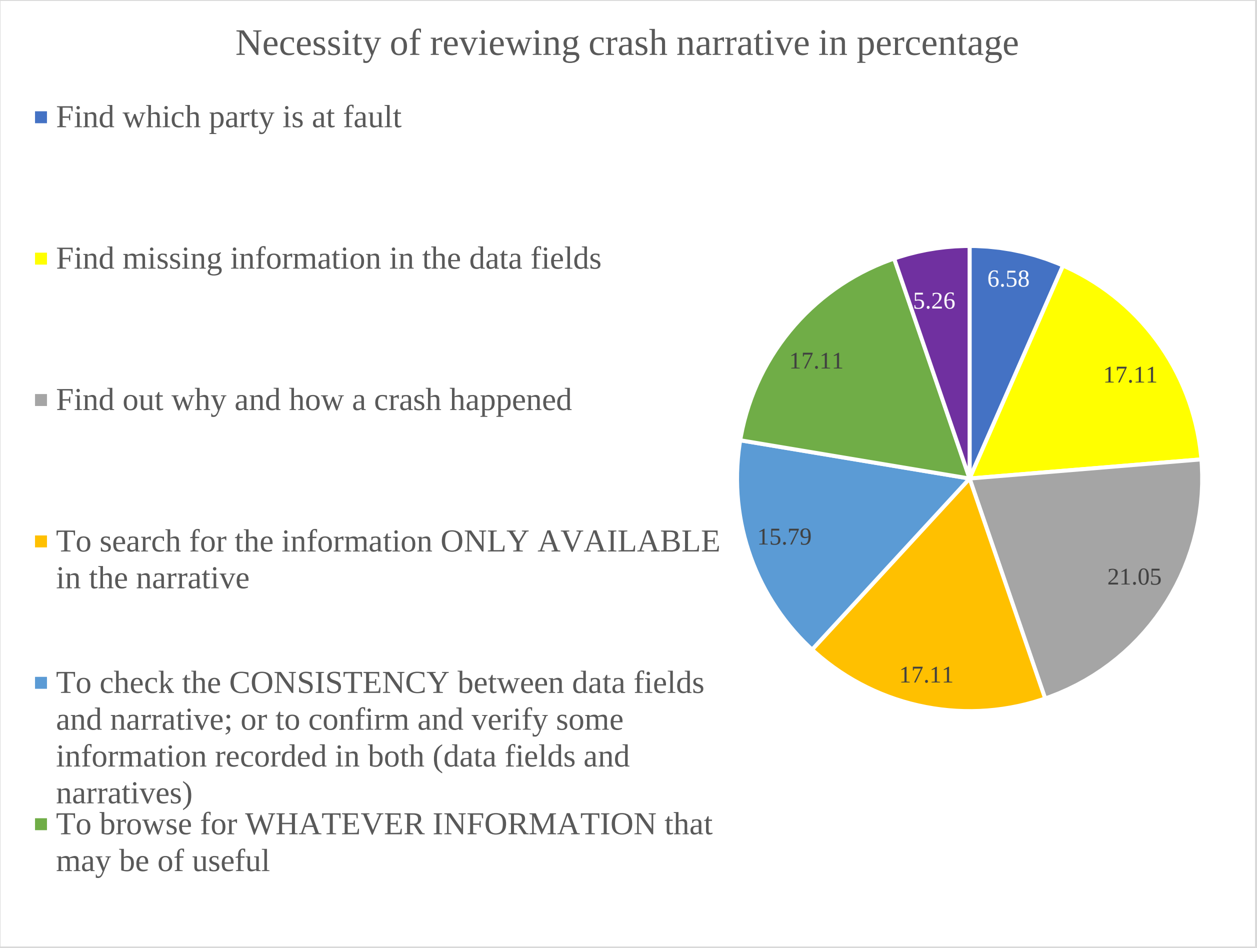

Figure 3-2 Survey Result: Necessity of Reviewing Crash Narratives

Other accounted for 4 of the 76 total selected responses. Those responses included that they only look at the narrative rather than the data fields (1), to confirm details of the crash (2), and to determine the actual location of the crash (1).

When asked “How often do you find the information you look for in the narrative?”. 87.5% of reh responders chose “sometimes”. Only 2 selected “always”. On average, it takes you an average of 2.56 minutes to complete reviewing the narrative section in one crash report, with minimum of 1 minute and maximum of 5 minutes. Considering there are over 100,000 reportable crashes took place every year in Wisconsin, the time spent on reviewing crash reports can be substantial. In particular, of the 16 respondents, all indicated that they use manual review to extract data from the crash narratives. Two also indicated that they also use some form of an automatic method.

The following **Error! Reference source not found.** shows the biggest challenges during your r eview of narrative section in a crash report? (check all that apply with a scale for the frequency of the challenges (in percentage) (5 being every time and 0 being Never)

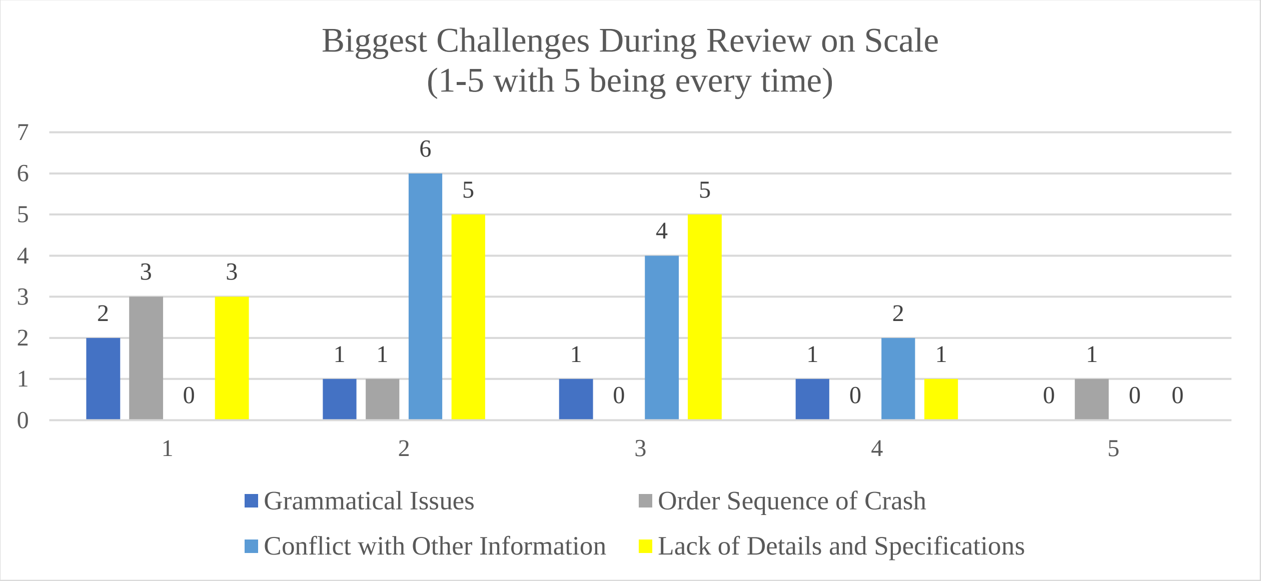

Figure 3-3 Survey Result: Challenge to Review Crash Narratives

A total of 43 selections were picked in response to the biggest challenge during the review of the narrative section in a crash report. 14 (32.56%) indicated that a lack of details and specifications was the most selected answer. Of those who indicated other, 3 indicated that the biggest challenge was regarding the location specifics. This could be the actual location of the crash to confirming the roadway identifiers. Two indicated that too much or two little information makes the narratives difficult as well. When rating the overall quality of the crash narratives, the mean of the recorded data is 3.63 out of 5 with 5 being very good.

The following Table 3-1 shows the survey responders' opinions on how to make crash narrative review more efficiently? (Check all that apply)

Table 3-1 Survey Result: Way to Make Narrative Review More Efficient

Methods to increase crash narrative review more efficient	**%**	**Count by 16 Responders**
Use text processing techniques (e.g. text mining and natural language processing) to help you screen crash narratives for key words and information before your review	27.27%	6

Use information retrieval and deep learning techniques to help you select relevant crash narratives for review	18.18%	4
If any information can be automatically extracted, what would you like to see?	9.09%	2
Other(please mention below)	22.73%	5
No	22.73%	5
Total	100%	22

If the respondent indicated the field of “If any information can be automatically extracted, what would you like to see?”, they were prompted to indicate what they would like to see. One respondent indicated that a stand-alone pdf of the diagram & narrative could help the data user. The other indicated that if possible, identifying if a vehicle crosses the roadway centerline, if a rear-end is due to queues from a signal or intersection, and if a vehicle were attempting to pass would be helpful.

Two respondents who indicated other made notes about intersection crashes to be drawn as a collision diagram with some standard details to be set. The following note was indicated regarding text processing and deep learning,

> “*Regarding text processing and deep learning, I believe both could be useful - but how useful would depend on how good those algorithms are. There are subtle clues often found in the narratives - both the picture that is drawn and what is written. Both of these offer insights into the crash. For instance, a severe or fatal crash will often have very detailed narratives and diagrams - which is an indicator that it was a noteworthy crash. Suffice to say, I think it would be challenging, but not impossible, to make this information for efficiently available. Since the diagram and narrative are so valuable, the biggest efficiency I see would be to make the diagram and narrative more accessible (i.e., withdrawn from the hardcopy on a separate stand-a-lone pdf). Then, between the electronic data and the "diagram/narrative" we would have all of the information we need to review crashes without asking for the "hardcopies" with the personal information drivers provide. So if a user could pull electronic crash data from the WisTransPortal, via the CMAA map for instance, and then be able to download "diagram/narrative" pdfs for all of the data pulled it would eliminate the need to request hardcopies and because no personal information would be exchanged - and I would think everyone could have access to it*”.

When asked “do you think the crash narrative can be improved?” Of those who responded yes, 6 commented on consistency of the reports, 5 on the details of the report, and 2 indicating that better location indication would be better. Comments on the consistency included adding set questions and added training. Improving details included always including specific information regarding why and how the crash occurred. Ideas for location included adding GPS coordinates for the crashes location and a diagram that shows where the vehicles began, collided, and ended. Many had overlaps regarding the consistency, details, and location details being things to improve upon.

In summary, a total of 16 responses were collected for this survey comprised of 14 data users and 2 data user & collectors. 93.75% of the data collectors identified themselves as consultants or government employees. The data users use the crash narratives for a variety of purposes, including Highway Safety Improvement Program (15), safety review for improvement programs (14), citizen requests (10), and traffic impact analysis (9).

All indicated that they use the narrative to look for detailed information and the rate that they find the information they needed on a scale of 1-5 was an average of 3.38. Most commonly, the narrative was used to find out why and how a crash happened (16), to search for information only in the narrative (13), to browse for whatever information may be useful (13), find missing information in the data fields (13), and to check for consistency (12). When asked how often the information is found within the narrative, the conclusion was sometimes.

Discussing the challenges of reviewing the crash narrative, the lack of details and specifications (14) and conflict with other information (12) were the largest problems. When asked the frequency of these issues on a scale of 0 – 5 (with 5 being every time), for the lack of details and specifications and conflict with other information the mean of the ratings were 2.29 and 2.667, respectively. To improve the review of the crash narrative, using text processing techniques was indicates 6 times. Other possibilities were identified as improving the consistency of the narrative by adding questions or required fields within the narrative and adding a GPS location of the crash.

Users want to see a more consistent narrative written. They do not believe that all narratives are hard to analyze but when too much or too little information is provided, they cannot find what they are looking for. To improve the narrative, I propose adding GPS coordinates to the crash location and specified questions to the crash narrative to ensure that all necessary details of the crash are reported.

The survey sample that was submitted by the data collectors (2) is too small to make conclusions. Most answers submitted were identified as not applicable to the questions. It was identified that intelligent work features would be helpful, or they would be helpful but might introduce some nuisances to the data collection process (e.g., similar to the function of MS word auto-spelling check).

4. METHODOLOGY

This section describes the principles and procedures used in the method for identifying missed crashes. To reduce manual work and speed up the process of identifying missed crashes, several machine learning techniques for text classification are considered in order to determine the best classifier. In this study, we implemented (1) multinomial naive bayes (MNB), (2) logistic regression (LGR), (3) Support Vector Machine (SVM), (4) Random Forest (RF), (5) K-nearest neighbor (K-NN), (6) recurrent neural network with Gated Recurrent Unit (GRU), and (7) a probabilistic classifier that combined probabilities using Noisy-OR. To the best of our knowledge, the Noisy-OR method has never been used in highway safety research. The rest of the section provides a summary of the seven methods. We also developed cascade classifiers using Noisy-OR and discussed the performance of the classifiers. Each method requires annotated training data (crash narratives); one as positive sample and the other as negative sample. The methodology to automatically obtain this training data is described later in the Case Studies section.

4.1 Multinomial Naive Bayes (MNB)

MNB is a classical text mining technique that is used in document classification. In MNB, the narrative is treated as a set of words. It uses a fixed set of words to define input vector, where the values in the vectors represent word frequencies in the narratives. The probability that a narrative indicates WZ crash is calculated by combining the prior probability of a narrative to be in WZ class with the conditional probabilities of words given that a narrative is in WZ class. The conditional probabilities are estimated by a smoothed version of maximum likelihood estimation that uses relative frequency counting. We applied Laplace smoothing to calculate relative frequency. More details on MNB can be found in (Manning, Schütze, and Raghavan 2008).

4.2 Logistic Regression (LGR)

LGR is a supervised linear classification algorithm which models the narratives using a logistic function called sigmoid function (Kantardzic 2011). It takes real numbers (i.e. features) as input and provides outputs between 0 and 1; and predicts the odds of being a narrative WZ based on the values of the independent variables. In our study, the independent variable represents the weight (i.e. count frequency, tf-idf) of the words. We applied L2 penalization in objective function to handle multicollinearity and overfitting problems (Zhang et al. 2019).

4.3 Support Vector Machine (SVM)

SVM was developed by Vapnik and his colleagues (Boser, Vapnik, and Guyon 1992; Cortes and Vapnik 1995) based on the principle of structural risk minimization in statistical learning theory. SVM has been found to be effective in many text classification problems such as hazard analysis

(Zhong et al. 2020), news article categorization and sentiment prediction (e.g., Joachims, 1998; Pang et al., 2002). It has been claimed to be less prone to overfitting (Joachims, 1998). As a supervised learning method that can be used for regression and classification, SVM uses kernel to map data in low dimensional space to a higher dimensional feature space and generates a hyperplane to separate the data by class. In this study, we used words as features and their number of occurrences as feature values. We applied linear classifier kernels because of its common use in text mining given that the word-based feature space is already very high dimensional.

4.4 K-Nearest Neighbor (K-NN)

K-NN is a non-parametric lazy machine learning algorithm used in the field of pattern recognition. It is one of the most widely used data mining techniques in classification problems. The classification score is calculated based on the majority votes of the k nearest narratives where the nearness is computed using a suitable distance measure. In our study, there are only two classes - WZ and Non-WZ or NWZ. The performance of K-NN depends on the distance measure used. Therefore, an appropriate distance measure must be selected to achieve the best K-NN performance. The two most common distance measures in the text classification field are Euclidean distance and cosine distance. We applied Euclidian distance metric, which is an extension of Pythagoras's theorem in multi-dimensional space. It calculates the distance by taking the square root of the sum of the squares of the difference between two narrative vectors and the value ranges from 0 to any positive number. We applied different values of K (i.e. 3, 5, 7, 9) for K-NN and found that K =7 gave the best result. More details of the K-NN can be found in (Cunningham and Delany 2020).

4.5 Random Forest (RF)

Random Forest was first developed by Tin Kam Ho (Ho 1995) based the random subspace method. RF is a learning algorithm based on ensembles of decision trees. RF is very popular in the field of pattern recognition and machine learning, and is used to solve high-dimensional classification problems (Breiman 2001). It fits several decision tree classifiers on various randomly selected sub samples (drawn with replacement) from the training set and takes the average of all probability predictions of the trees to improve accuracy and control overfitting (Breiman 1999).

4.6 Gated Recurrent Unit (GRU)

Recurrent neural network (RNN) is a special neural network architecture for handling sequential data such as text narratives which are sequences of words. While processing sequential inputs one item at a time (for example, one word at a time), RNN needs a mechanism to learn to remember important items it saw earlier and forget the unimportant items. This is achieved using special neural networks cells. The two most used such cells are Gated Recurrent Unit (GRU) and long short-term memory (LSTM). The GRU proposed by Cho et al. (Cho et al. 2014) is similar to LSTM with a forget gate (Gers and Cummins 1999); but compared with traditional LSTM, it has a shorter training time and fewer parameters. Due to these advantages, we used GRU in this work. The GRU does not have any cell state like LSTM and uses hidden state to transfer information. The GRU has two gates: the reset gate decides how much past information will be transferred to the next step, and the update gate decides what information to be added or discarded to the current layer. The update gate is very similar to the forget and input gate of an LSTM. Readers are referred to (Cho et al. 2014) for more information.

4.7 Noisy-OR Based Classification

In the Noisy-OR method, the probability of being a specific type of crash narrative is calculated by combining the probability scores of unigrams (words) and bigrams (two consecutive words) in the narrative. It is a probabilistic extension of logical "or" (Ońisko, Druzdzel, and Wasyluk 2001; Vomlel 2006). If any input has a high probability score (such as a value close to1) then the combined probability in Noisy-OR becomes high. The combined probability in Noisy-OR is even higher if more input probabilities are high.

To apply Noisy-OR classifier to crash narratives, we need to compute the probabilities of unigram, bigram, and trigram and combining these probabilities using the Noisy-OR method, which are discussed in the following section.

4.7.1 Equation of Simple Count Probability

For every unigram, bigram, and trigram *w* in the corpus, the method first computes the probability that if it is present in a narrative, then the narrative is positive, i.e., P(positive|w). Then, this probability is computed using simple frequency counts, as shown in Equation 1.

$$Probability\ Score\ (w) = \frac{\text{Positive Count(w)+1}}{\text{Positive Count (w)+Negative Count(w)+2}} \qquad (1)$$

where w is a unigram, a bigram, or a trigram. *Positive Count* means the number of occurrences of w in the positive narratives. Similarly, the *Negative Count* indicates the number of events of w in the negative narratives.

The Equation essentially computes out of all the narratives in which w occurs how many narratives are positives, which is the probability that a narrative will be positive if w occurs in it. Then, smoothing is applied by adding one in the numerator and two in the denominator of the Equation. This simple version of Laplace smoothing assumes w occurred at least once in a positive narrative and a negative narrative. Smoothing done in this way ensures that among the unigrams, bigrams, and trigrams that have zero negative counts, the ones with higher positive counts receive higher probability scores. Otherwise, they will all receive an unrealistic probability score of 1 because they occurred in a few positive narratives and no negative narratives.

In case of the words that appear in both positive and negative narratives with very high frequency (Count), it is likely to reduce the probability of that specific word. For example, if a unigram '*unit*' appears in the narratives of a specific type of crash (positive case) 110,933 times and in all the other narratives (negative cases) that excludes that specific crash 1,000,904 times, which according to Equation 1, gives a probability of 0.099. It indicates that the word is not relevant for the classification task. On the other hand, if a unigram/ bigram/ trigram appears in both positive and negative narratives with high frequency but has a higher frequency in positive narratives, Equation 1 gives a good probability score to the corresponding unigram/ bigram/ trigram. For example, if a unigram '*inattentive*' appears in the narratives of a specific crash 2743 times and in all the other narratives 1808 times, which according to Equation 1, gives a probability of 0.6023, indicating that the word is relevant for the classification task.

To classify a given narrative as positive or negative, its probability of being positive is computed by combining the probability scores of the unigrams, bigrams, and trigrams present in it. The method needs to compute P(positive|w_1,w_2,...,w_n), where w_1..w_n are unigrams, bigrams, and trigrams present in the narrative. It computes it by combining the probabilities P(positive|w1), P(positive|w2),…, P(positive|wn), which have been computed as described earlier. Noisy-OR is a method of combining probabilities (Zagorecki and Druzdzel 2004), which is commonly used in Bayesian networks (Oniśko et al. 2001; Vomlel 2006). Instead of true/false values in Noisy-OR, the inputs and output are probabilities (hence termed "noisy"). Analogous to logical "or", in Noisy-OR, if any one of the input probabilities is high (i.e., close to 1), then the combined probability is high. But unlike logical "or", the combined probability is even higher if more input probabilities are high. The combined probability is low (i.e., close to 0) only when all the input probabilities are low. Noisy-OR combined probability is mathematically computed as shown in Equation 2, where the probability score of a narrative is calculated by combining the probability scores of unigrams, bigrams, or trigrams occurring in it.

$$Noisy - OR\ Proability\ Score\ (N) = 1 - \prod_{i,j=1}^{n}(1 - P_i)^j \qquad (2)$$

where N is a given narrative, P_i indicates the probability score of ith unigram, bigram, or trigram as computed from the training data using Equation 1, and j means the number of occurrences of that ith unigram, or bigram or trigram in the crash narrative N.

It should be clear from Equation 2 that if there is no unigram, or bigram, or trigram in a narrative with a high probability score, then the probability score of the narrative will be close to zero. On the other hand, a single unigram, or bigram, or trigram with a high probability score will result in a high probability score of the entire narrative. This fact is precisely the behavior that has been observed in our data. Furthermore, more unigrams, bigrams, and trigrams with high probability scores make the combined probability score higher.

4.7.2 Equation of Weighted Count Probability

The probability scores computed using Equation 1 will be adversely affected if the number of negative narratives is disproportionately higher than the number of positive narratives. In the Weighted Count Probability Equation, the positive counts are weighted by the average number of positive word appearance in the positive narratives. It is designed to capture the unigrams/bigrams/trigrams that appear not only more often in positive narratives than negative narratives, but also more often per positive narrative than per negative narrative. The Weighted Count Probability is formulated in Equation 3:

$$Probability\ Score\ (w) = \frac{\text{Positive Count(w)}*\left(\frac{\text{Positive Count(w)}}{\text{Number of instances in positive}}\right)+1}{\left(\text{Positive Count (w)}*\left(\frac{\text{Positive Count(w)}}{\text{Number of instances in positive}}\right)\right)+\left(\text{Negative Count(w)}*\left(\frac{\text{Negative Count(w)}}{\text{Number of instances in negative}}\right)\right)+2} \tag{3}$$

Here, *Positive Count* and *Negative Count* represent the same meaning as in Equation 1. *Number of instances in positive* means the total number of reported cases for distracted or inattentive. The *Number of instances in negative* means the total number of cases not reported as distracted or inattentive.

With the probability scores obtained using Equation 3, the probability of a positive narrative is computed using the Noisy-OR method described earlier. Given that a positive narrative will have an indicative word mentioned more than once, the Noisy-OR probability score of the narrative will increase accordingly (note that in Equation 2, $(1 - P_i)$ is raised to the power of j, the number of occurrences). In contrast, a negative narrative that has fewer indicative words will have a lower probability of being positive.

4.7.3 Model Development for Three Classes Classification

If a crash is potentially associated with two or more types, classifying the crash based on its narrative can be challenging. In this study, distinguishing between distracted driving (DD) related crashes and inattentive driving (ID) related crashes exemplifies this challenge as these crashes, also known as distracted or inattentive (DOI), may not be adequately flagged by the police officers who filled out the crash report form, and there could be no data field to differentiate between DD and ID. Our strategy is to extract pertinent information that is more relevant to one type than the other if such a dominance exists. Accordingly, two types of models have been developed: the hierarchical model and the priority model.

Figure 4-1 shows the hierarchical model where the Noisy-OR score by the DOI classifier for a specific narrative pass through the DD and ID classifiers, respectively, and result in separate Noisy-OR scores S1 and S2. This model can first identify NDOI (non-distracted and/or inattentive) cases if the score of the DOI classifier is lower than a threshold value (shown as “thrs” in Figure 4-1). Next, if S1 is greater than S2, the narrative is classified as a DD narrative; otherwise, an ID narrative.

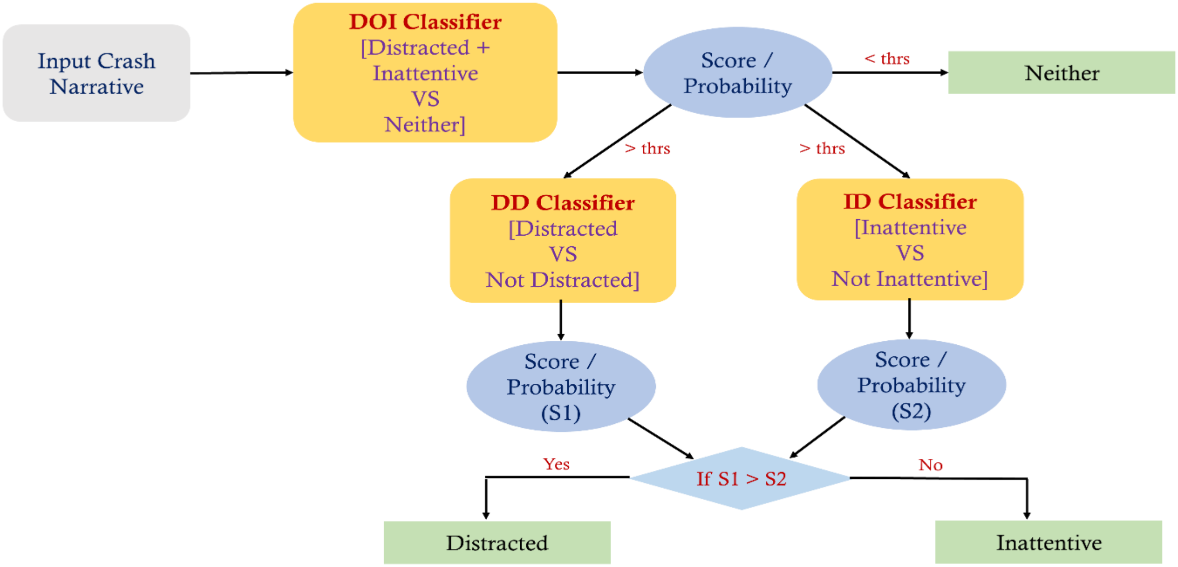

Figure 4-1 Hierarchical Model

However, the model performance can be susceptible to the imbalanced performance of the DD and ID classifiers. For example, the DD classifier may perform much better than the ID classifier if the DD classifier has more extensive training dataset than the ID classifier and/or if the DD classifier has more distinctive unigrams and bigrams than the ID classifier. The limitation leads to the creation of models based on priority.

Figure 4-2 and Figure 4-3 are both priority models in which Figure 4-2 gives high priority to the DD classifier while Figure 4-3 assigns high priority to the ID classifier. In the priority model-DD (Figure 4-2), the crash narrative passes through the DOI classifier first. This classifier will determine if the given narrative is DOI or NDOI depending on a threshold value (thrs1 in Figure 4-2). Next, if the Noisy-OR score by the DOI classifier is greater than the threshold (thrs1), the narrative will go through the DD classifier. The purpose of the DD classifier is to determine if the given narrative is a “distracted” narrative or not. If the generated Noisy-OR score by the DD classifier is greater than the threshold (thrs2 in Figure 4-2), the narrative is classified as a “distracted” narrative; otherwise, it will go through the ID classifier. Then, if the Noisy-OR score by the ID classifier is greater than the threshold (thrs3 in Figure 4-2), the narrative is classified as an “inattentive” narrative; otherwise, it will be classified as NDOI narrative.

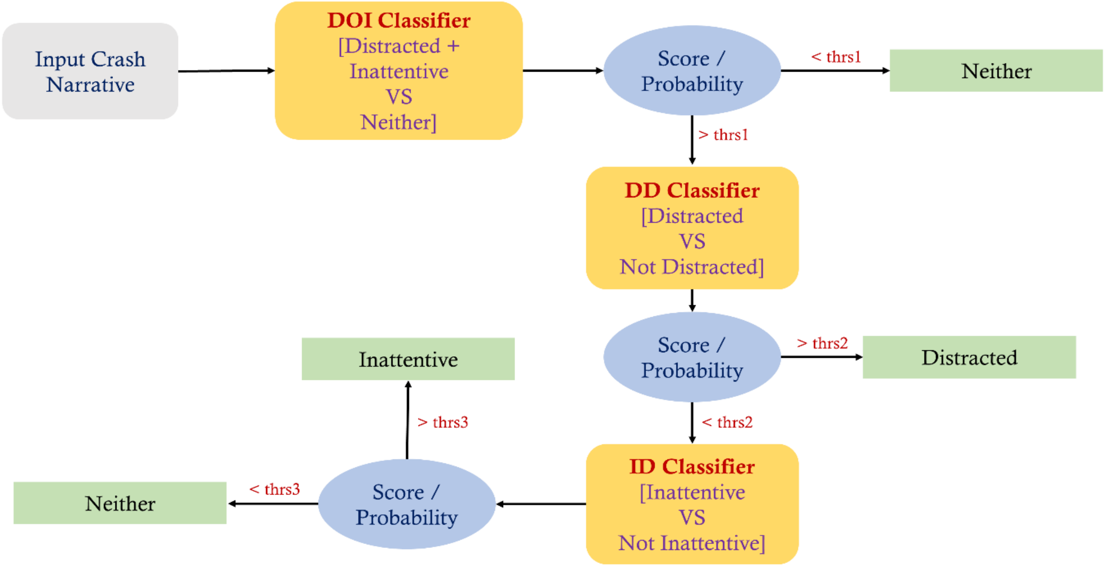

Figure 4-2 Priority Model-DD Classifier having Higher Priority

In the priority model-ID (Figure 4-3), the crash narrative first passes through the DOI classifier, where a threshold value (thrs1 in Figure 4-3) determines if the given narrative is DOI or NDOI. Next, if the Noisy-OR score by the DOI classifier is greater than the threshold (thrs1), the narrative will go through the ID classifier. If the generated Noisy-OR score by the ID classifier is greater than the threshold (thrs2 in Figure 4-3), the narrative is classified as an “inattentive” narrative; otherwise, it will go through the DD classifier next. Then, if the generated Noisy-OR score by the DD classifier is greater than the threshold (thrs3 in Figure 4-3), the narrative is classified as a “distracted” narrative; otherwise, it will be classified as an NDOI narrative.

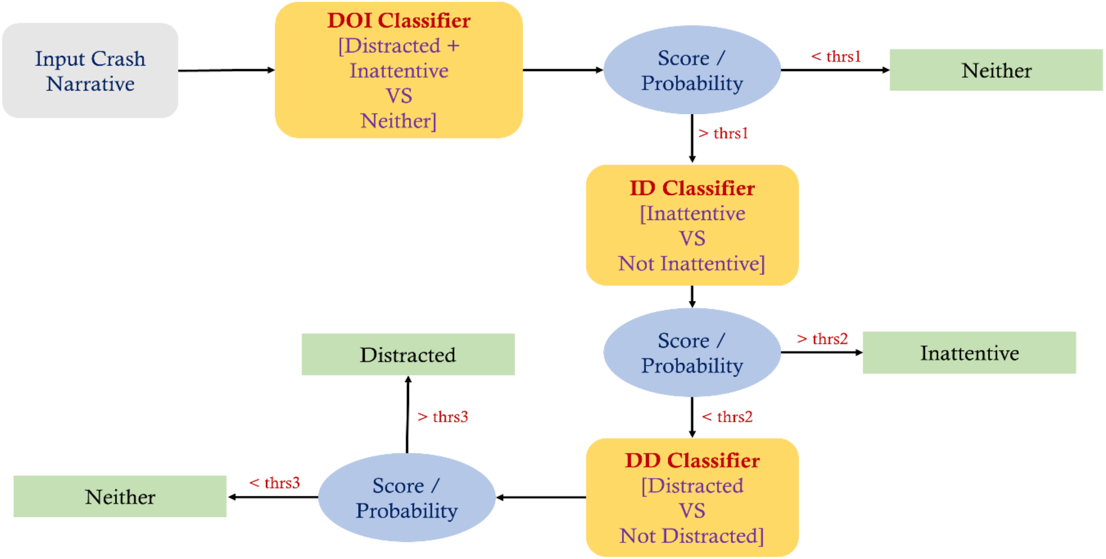

Figure 4-3 Priority Model-ID Classifier having Higher Priority

5. CASE STUDIES

In this section, two case studies are conducted with all seven methods introduced in the Methodology section. One is to identify unaccounted work zone crashes based on crash narrative and the other is to find missed distracted/inattentive driving related crashes from crash narrative. In the second case study, distracted driving crashes are further separated from inattentive driving cases. All the model results are summarized and compared; their performance is discussed; and the most appropriate ones are recommended.

5.1 Work Zone Crashes

Work zone activities are essential for maintaining good roadways, supporting economic development and competition, and improving safety. While road work is temporary, the poor decisions and mistakes made by motorists that lead to work zone crashes can have lasting impacts. According to the Federal Highway Administration (FHWA), 27,037 people, or 773 per year, died in work zone crashes in the U.S. from 1982 through 2017 (CDC, 2020). In Wisconsin, more than 2,600 work zone crashes took place every year over the past five years, resulting in 5,200 injuries and 50 deaths (WisDOT, 2020). Work zone safety for both motorists and workers is an urgent issue that must be addressed through better design, operations and management. Work zones near traffic, whether they involve major road construction, utility work, or emergency vehicles at the side of the road, always present some risk to both drivers and workers. Identifying and analyzing historical work zone crashes can save lives.

Observational safety analysis has been instrumental in identifying potential deficiencies in work zone design and traffic operations. Examples of safety analyses based on crash data include: crash rate estimation across different work zone configurations (Cheng et al. 2012; Daniel, Dixon, and Jared 2000; Elias and Herbsman 2000; Khattak, Khattak, and Council 2002); crash pattern identification and categorization(Garber and Zhao 2002; Graham, Paulsen, and Glennon 1978; Weng et al. 2016); work zone crash prediction (Li and Bai 2009b; Meng, Weng, and Qu 2010); and evaluating the safety of innovative work zone designs and management strategies (Li and Bai 2009a; Maze, Burchett, and Hochstein 2005; Rahman et al. 2017; Ullman et al. 2008). All the aforementioned examples are dependent on the completeness and accuracy of work zone crash data. The crash in the structured data may not have been coded or recorded as that specific crash type.

5.1.1 Data Collection

The dataset comprised 377,479 crash reports that occurred between January 1, 2017 and October 31, 2019 that were acquired from the Wisconsin Department of Transportation (WisDOT) through the WisTransPortal data hub. A construction zone flag (CONSZONE) within the crash data indicates whether "*a crash occurred in a construction, maintenance, or utility work zone or* *is related to activity within a work zone*". The reported work-zone (WZ) crashes make up 2.27%, 2.49%, and 1.93% of total crashes for years 2017, 2018 and 2019, respectively. Narratives were included in 94.21% of the reported WZ crashes and 77% of the non-work zone (NWZ) crashes. The ratio of WZ to NWZ crashes is 1:36, which is a highly imbalanced dataset.

The two following sample crash narratives were randomly chosen from the dataset to illustrate the structure of crash narratives.

WZ crash narrative example: *"Entering construction zone with right lane closure. Unit 1 driver stated unit 2 and a semi were straddling center line. Unit 1 driver stated thought unit two was merging to right lane toward hwy c exit and tried to pass unit 2. Unit 1 driver stated himself and semi were straddling traffic lane to stop other drivers from passing on right as right lane was closed ahead. Unit 2 stated unit 1 attempted to pass on left shoulder but ran out of room due to portable warning sign. Unit 2 driver stated unit 1 driver side swiped driver side."*

NWZ crash narrative example: *"Unit #2 was stopped in the inside straight lane of eastbound university ave., at a red light at the intersection with n. Midvale blvd. Unit #1 was traveling in the same lane directly behind unit #2, and was unable to stop in time to avoid a rear end collision with unit #2. The roadway was wet, and the weather conditions were rainy."*

The numeric values within the narratives usually represent date, time, driver and road information. The narratives have a certain formality but can still be flexible in the sequence of events. In the WZ narrative, some sentences contain words that indicate WZ (e.g., "construction zone", "right lane closure", "portable warning sign"), while others do not contain any WZ indicators. In fact, the latter cannot be distinguished from sentences that could have been in a NWZ narrative. This observation is true of other WZ narratives as well; only a few words are indicative of a WZ while the rest of the narrative is not, suggesting that presence of just a few words can be used to identify a WZ narrative without having a deep understanding of the entire narrative. Additionally, there are no such words in the narrative that specifically indicate NWZ.

In this study, the 2017 and 2018 work zone crash data were used to train a classifier (described later) to categorize a narrative as either WZ or NWZ and the NWZ narratives of 2019 (Data was available till October 31, 2019) were used as test data to recover missed WZ crashes. The narratives corresponding to reported WZ crashes (i.e., marked under CONSZONE flag) were used as examples of WZ narratives to train the classifier. Similarly, the narratives corresponding to reported NWZ crashes (i.e., not marked under CONSZONE flag) were used as examples of NWZ narratives. The method did not require the manual annotation of training examples, a task that usually requires the huge effort of training a classifier. However, the training dataset created does include a high level of noise. On one hand, many narratives of reported WZ crashes may not have any relevant information about the WZ. For example, the officer may have already indicated a crash as WZ by using the CONSZONE flag, hence not feeling the need to mention it in the narrative. However, WZ crashes are known to be missed, and there are narratives corresponding to reported NWZ crashes that are actually WZ. The classifier may have difficulty learning from such noisy training data.

5.1.2 Data Cleaning and Pre-Processing

Several text mining techniques for data cleaning and pre-processing were applied to prepare the data. The key terminologies from the text mining domain are introduced here:

- *Corpus* is the collection of all the narratives.
- *Tokenization* is the process of breaking up the sentence into a token. A token can be words, numbers, a punctuation, unigram, or bigram. The terms unigram and bigram are used interchangeably as the token in this study.
- *Collection frequency* (cf) is the number of times a token occurred in the corpus.
- *Term frequency* (tf) is the number of times a token occurred in a narrative.
- *Document frequency* (df) is the number of documents/narratives that contain a token. Only the tokens with high df values in WZ narratives will have a high impact on the model.

In the training dataset, the narratives were first lower-cased to merge the occurrences of the same word in different cases. Then, all punctuations and special characters (e.g., ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [/] ^ _ ` { | } ~) were removed from the narratives. Next, the narratives were converted into tokens to build a vocabulary list from the training set. The narratives may include spelling errors and/or words in multiple forms, such as “zone” and “zones” or “construction” and “construct”, which are common issues when mining unstructured text data. While some text mining techniques can handle these issues, there is no guarantee the problem will be solved completely. Furthermore, improper processing of these words may lead to new problems. Thus, the words in the vocabulary list were kept as-is.

Research shows that machine learning algorithms cannot provide good performance for an imbalanced dataset that has far fewer number of examples of one class compared to the other (Jeong et al. 2018). Based on the ratio of reported WZ and NWZ crash narratives, our data set can be called an unbalanced dataset (Leevy et al. 2018). A balanced dataset can be obtained by oversampling or undersampling. For some classifiers (such as SVM), oversampling can degrade their performances (Glen 2019). To create a balanced dataset for MNB, LGR, SVM, RF, K-NN and GRU, we randomly selected 2,000 WZ crash narratives and 2,000 NWZ crash narratives from crash reports from the years 2017 to 2018, resulting in the number of feature vectors in training set to be 4,000. Similarly, we chose another 4,000 as validation data from crash reports from the years 2017 to 2018. The crash narratives from the year 2019 were used for evaluation which is later described in Section 5.

Simple data processing techniques, such as case folding (turning words into lowercase), punctuation and word spacing removal techniques are applied to prepare data for all methods. In addition, all redundant terms (such as stop words,) and the words with length (number of characters) less than 4 were deleted from the narratives for the methods GRU, LGR, SVM, RF, and K-NN models.

5.1.3 Feature Generation and Model Parameters Tuning

The tasks of feature generation and model parameter tuning are different among the classifiers, which are described sequentially in this subsection. In *Google Colab*[1], we used python as programming language, the machine learning library *TensorFlow*[2] for GRU, and the machine learning library *sklearn*[1] for MNB, LGR, SVM, K-NN and RF to generate features and develop models. After processing the narratives, we converted narratives into tokens (unigrams) by count vectorization. After trying input vectors with various lengths such as 50, 100, 200, 300, 500, 1000, 5000, and the full length of vocabulary, we found that the input vector with length 500 gives the best result for MNB, LGR, SVM, K-NN and RF in the reported WZ in the training dataset. In this process, we built a vocabulary that only considers the top 500 words ordered by word frequency across the narratives. We fine-tuned other model-specific parameters of all models based on training dataset and used the best parameters for evaluation. We also tested advance data processing techniques such as lemmatization, bigram tokenization, tf-idf weighting, and different vectorization architectures to train MNB, LGR, SVM, K-NN, and RF. No significant improvement was observed, rather the model performance degraded for some.

For GRU, we used 154, the third standard deviation of the narrative length, as the input vector length and applied post padding to the vector to fill with 0 if the input length was shorter. The tokens were converted to vectors using pre-trained Google *word2vector*[3]. Words that did not exist in the dictionary were initialized with a random number in the range of 0 to √0.25 using Gaussian distribution. We developed GRU by stacking two GRU layers (each containing 32 hidden units) and a Dense layer (containing 1 hidden unit with sigmoid function). The Dense layer provides the final output of the model. We used binary cross-entropy as the loss function, 32 as the mini-batch size, Adam as the optimizer, and the early stopping in the callback function to find the best model.

We implemented Noisy-OR in python. After data processing, unigrams and bigrams were extracted from all WZ and NWZ narratives of 2017 to 2018. Then, the probability score of each

[1] Google Colab at https://colab.research.google.com/notebooks/intro.ipynb#recent=true
[2] TensorFlow at https://www.tensorflow.org/api_docs/python/tf/keras/layers/GRU
[3] Google wor2vec was downloaded from https://code.google.com/archive/p/word2vec

unigram and bigram was calculated using Equation (1) to be used in the Noisy-OR method. Unigrams and bigrams with probability scores less than 0.25 or those that appeared less than four times in the narratives were discarded.

5.1.4 Results and Analysis

In this study, we have implemented 7 classifiers: Noisy-OR, MNB, LGR, SVM, K-NN, RF, and GRU. We used n-grams (e.g., unigrams and unigram+bigram) with Noisy-OR, which are discussed in this section. The characteristics of missed crashes are analyzed from spatial and temporal perspectives, along with other features. The additional analysis is expected to provide insight on the circumstances in which crashes are not reported as WZ related so that recommendations can be made for improving future data collection.

During our manual review for evaluation purpose, we observed that the narrative could have multiple WZ-related keywords that would help identify missed WZ crashes. Therefore, analyzing the keywords captured by a classifier during the training phase can provide insights about the classifier. The 2017-2018 crash data were cleaned and preprocessed, showing 10,875 unigram and 96,550 bigram words (tokens) in the corpus. Table 5-1 presents the top ten positive unigrams and bigrams and their corresponding probability scores. As shown in Table 5-1, the bigram approach extracted more WZ-related information than the unigram approach. However, despite high probability scores, some positive unigrams did not carry meaningful information such as “Kampo”, “Kucej”, or “Werych”. While “Kampo”, “Kucej”, and “Werych” may appear only in WZ cases, at a very low frequency, meaning including them in the Positive Unigram list may degrade the model’s performance. For example, if a narrative has many such unigrams, the Noisy-OR may tend to classify it as a WZ crash even if it’s not.

The document frequency (df) and collection frequency (cf) of the training set were calculated to examine how the positive unigrams and bigrams with high probability scores influence the proposed method. The classifier performance did not degrade much due to lower document frequency(df) of the less meaningful positive unigrams and bigrams. Thus, an important positive token should have both high df and cf values and with high probability score.

Table 5-1 Top Ten Positive Unigrams and Bigrams by Probability Score Using Equation 1

Rank	Positive Unigram		Positive Bigram	
	Positive Words	**Probability**	**Positive Words**	**Probability**
1	flagman	0.960	active construction	0.990
2	taper	0.947	in construction	0.988

3	barreled	0.937	temporary cement	0.983
4	dividers	0.929	zone where	0.972
5	roadworks	0.923	construction crew	0.971
6	kampo	0.917	zone lane	0.964
7	unfinished	0.917	interstate is	0.960
8	flaggers	0.917	no workers	0.960
9	kucej	0.909	flag person	0.957
10	werych	0.900	workers present	0.956

Table 5-2 populates a list of the top 15 important positive unigrams and bigrams ranked by df, cf and probability score (p_r) in a decreasing order. In the positive unigram list, the token “construction” is the most important because it has the highest df and cf values. The most important token in the positive bigram list is “construction zone”. Approximately 35.15 % of the WZ crash narratives contain the token “construction”, whereas 16.16% of WZ crash narratives contain “construction zone”. Table 5-2 shows that the positive bigram list offers more specific WZ crash information and higher probability scores than the unigram list.

Table 5-2 Top 15 Positive Unigrams and Positive Bigrams By df, cf, and Probability Score

Rank	Positive Unigram				Positive Bigram			
	Token	**cf**	**df**	**Pr**	**Token**	**cf**	**df**	**Pr**
1	construction	2960	2088	0.89	construction zone	966	826	0.9
2	zone	1181	972	0.45	the construction	763	625	0.82
3	closed	743	588	0.44	a construction	484	437	0.77
4	barrels	407	314	0.69	to construction	320	312	0.73
5	closure	265	191	0.61	was closed	242	228	0.51
6	orange	192	152	0.34	construction barrels	195	167	0.77
7	barrel	228	147	0.56	lane closed	212	161	0.67
8	temporary	170	126	0.37	construction unit	158	156	0.78
9	zoo	219	123	0.56	under construction	151	149	0.79
10	cones	166	122	0.45	construction area	158	136	0.82
11	workers	120	110	0.52	road construction	145	135	0.68

12	barriers	119	97	0.49	work zone	161	132	0.92
13	barricades	107	78	0.42	the zoo	206	120	0.67
14	attenuator	145	74	0.47	construction and	123	120	0.7
15	worker	95	67	0.51	zoo interchange	181	114	0.67

* cf = collection frequency in WZ narratives, df = document frequency in WZ narratives, p_r = probability.

The unigram method will classify a narrative as a WZ crash if the narrative contains the token “construction” (p_r= 0.89) from the positive unigram list only because the threshold value is greater than or equal to 0.89. The df of “construction” is much higher compared to other unigrams in the list, so the misclassification rate by the unigram method will be higher. Compared with the positive unigram “construction”, the positive bigram “construction zone” (p_r= 0.90) is more contextual and has a higher df than the remaining bigrams in the list. A narrative with the presence of “construction zone” instead of “construction” is more likely to be correctly classified as a WZ crash. The manual review result shows that all of the NWZ narratives that contain “construction zone” are true WZ crashes. However, 22 NWZ narratives that contain “construction” are not WZ crashes.

Positive tokens such as “fst”, “kampo” and “kicmol” in the positive unigram list do not carry any meaningful information. These unigrams have a small df with high probability scores, meaning they should be discarded to reduce the misclassification rate. The positive token lists also contain names of locations such as “zoo” in unigram and “the zoo[4]” in bigram. The presence of those tokens can cause the Noisy-OR method to misclassify NWZ crashes as WZ crashes.

The results of LGR (Zhang et al. 2019) and linear SVM (Chang and Lin 2008; Cuingnet et al. 2011; Guyon et al. 2013) can be explained by the magnitudes of the coefficients of the words. However, their interpretations are different. In linear SVM, if the absolute value of the coefficient of the vector component (word) is smaller than the coefficients of other components, the coefficient of this component has little influence on the classification result, and vice versa (Cuingnet et al. 2011). In LGR, the coefficient indicates the log odds of being a positive class (WZ crash), and the exponent of coefficient $(e^{coefficient})$ indicates the log odds ratio. For example, the log odds of WZ crash is 4.34 times higher when a narrative has the token “construction” (Figure 5-1) compared to the narratives that do not have this word. Or the odd of a narrative being a WZ crash is 76.71 times higher for the presence of word “construction” compared to not present of that word in the narrative. Due to difference in interpretations, the LGR cannot be directly compared with SVM using the coefficient of component (token). Figure 5-1 shows the top 30 words of LGR and SVM. There are 14 words with positive coefficients and 16 words with negative coefficients in LGR, and there are 17 words with positive coefficients and 12 words with negative coefficients.

[4] Zoo interchange construction is the most complex and expensive highway project in Wisconsin’s history, which began in 2014 with an expected completion date of 2022.

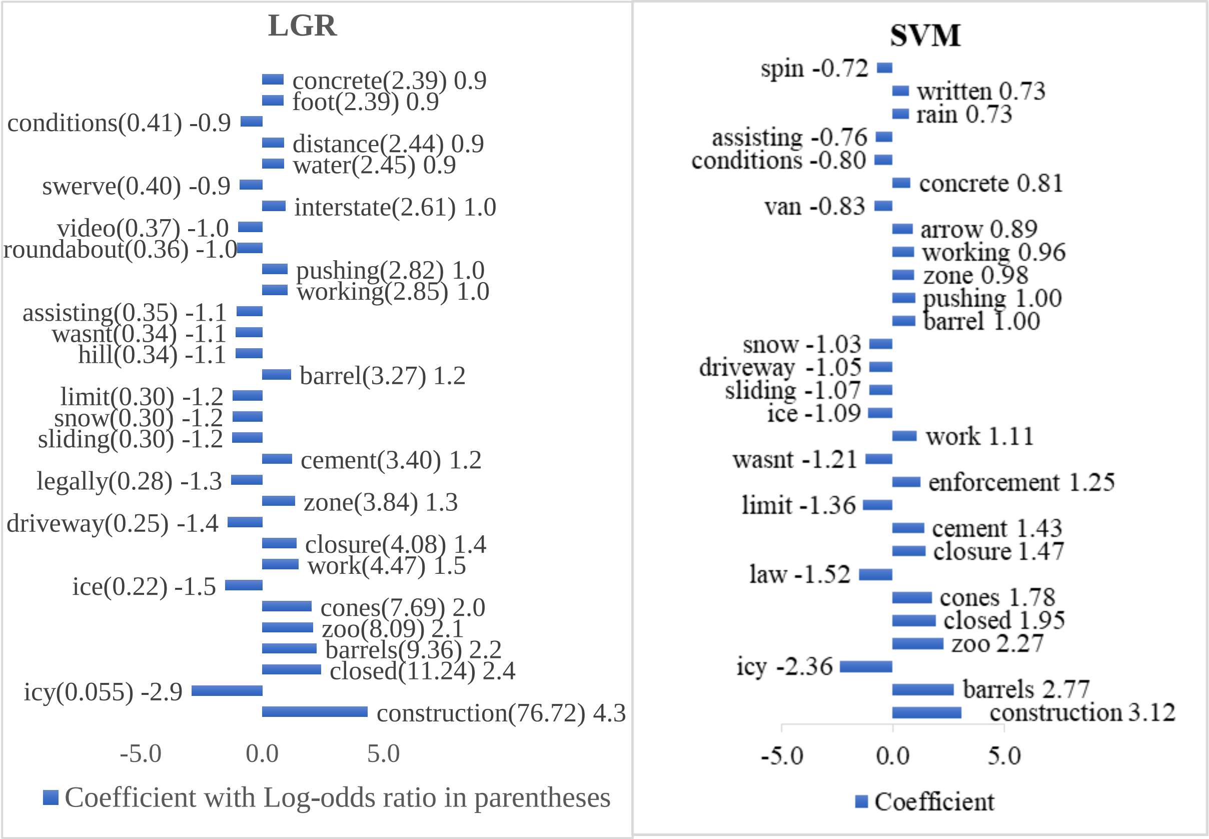

Figure 5-1 Important Words found by LGR and SVM

For Noisy-OR, we selected important words based on probability score (pr). Interestingly, the words in the top 15 unigrams and bigrams of the Noisy-OR word list (Table 5-2) are also common in the word lists of LGR and SVM (Figure 5-1). But, there are no words (unigram and bigram) with negetaive coefficient like LGR and SVM in Noisy-OR that can lead a classifier to conclude negative class. The number of positive unigrams and bigrams with probability scores greater than 0.25 are 154 and 1,665, respectively. Although LGR and SVM can identify good positive words (words with high coefficient values), the irrelevant words with negative coefficient actually degrades the performacne of the model.

The deep learning models (GRU in our case) are often regarded as blackbox models because the internal mechanism of the alogrithm is not interpretable. In this study, we gained some insights on our GRU model from its output. The manual review of GRU's top 100 narratives helped reveal important keywords. For example, the 100 narratives with top GRU scores contained the WZ related word "construction", which means that GRU emphasized this word while classifying the narratives. The other important words in the narratives are very similar to that of LGR, SVM, and Noisy-OR.

5.1.5 Model Evaluation and Discussion

For our first evaluation, the standard evaluation metric of area under ROC curve (AUC) was used to measure overall performance of all the classifiers. We randomly selected 100 WZ and 100 NWZ reported crash narratives from the 2019 crash reports, and manually marked the true positives (missed WZ crashes) and true negatives (true NWZ crashes). We found that only 36 cases were truly positive in WZ narratives, and all were true negatives in NWZ narratives. The surprising result of WZ narratives forced us to review another 200 WZ reported cases, and that time we found 56 true positives. In this way, we obtained 92 true WZ cases and 208 NWZ cases from the 300 reported WZ narratives; and 100 true NWZ cases from the 100 reported NWZ narratives. In total, the test dataset contained 92 WZ cases and 308 NWZ cases out of the total 400 crash reports. We used this manually reviewed dataset to compare the models using AUC. We can infer from the above numbers that the WZ crash narrative in the training data contains approximately 70% or 208/300 noisy narratives that do not contain any WZ related words. During the manual review process, we also observed that only a few important keywords or phrases in the true WZ narratives are relevant for classification.

From the results obtained using first evaluation described above, we found that Noisy-OR (unigram+bigram) achieved the highest AUC score (0.98) and GRU achieved the second highest AUC score (0.97). LGR (0.96) and SVM (0.95) provided similar AUC scores, while MNB and RF had AUC scores of 0.95 and 0.87, respectively. The K-NN achieved the lowest AUC score (0.65). We also found that the ROC curves of Noisy-OR, GRU, SVM and LGR follow similar trend. Since the differences in AUC values for these models are small, the AUC cannot be used to determine the best model. While constructing the test dataset through manual reviewing, we had found that it had many easy WZ and NWZ cases, which may be the reason for the small differences in AUC values.

Although the AUC method helps in evaluating the classification performance of the classifiers, it does not help in evaluating the performance of the classifiers in identifying missed WZ crashes. Therefore, we conducted a second evaluation which closely reflects the classifier’s ability to find missed WZ crashes from reported NWZ crash narratives. We ran all the classifiers on the 2019 NWZ narratives (total 82,215 crash reports that were flagged as NWZ by the “CONSZONE” flag) and collected the top 100 narratives of each classifier ranked by the classification scores assigned by the classifier. We then manually evaluated the top 100 narratives of each classifier. The classifier that includes the maximum number of missed WZ crashes in its top 100 narratives is deemed as the best classifier.

Figure 5-2 shows the accuracy of the seven classifiers for our second evaluation. As was described above, for each of the seven classifiers, we selected 100 narratives (7×100 = 700 narratives in total) that had the highest classification scores and manually evaluated them.

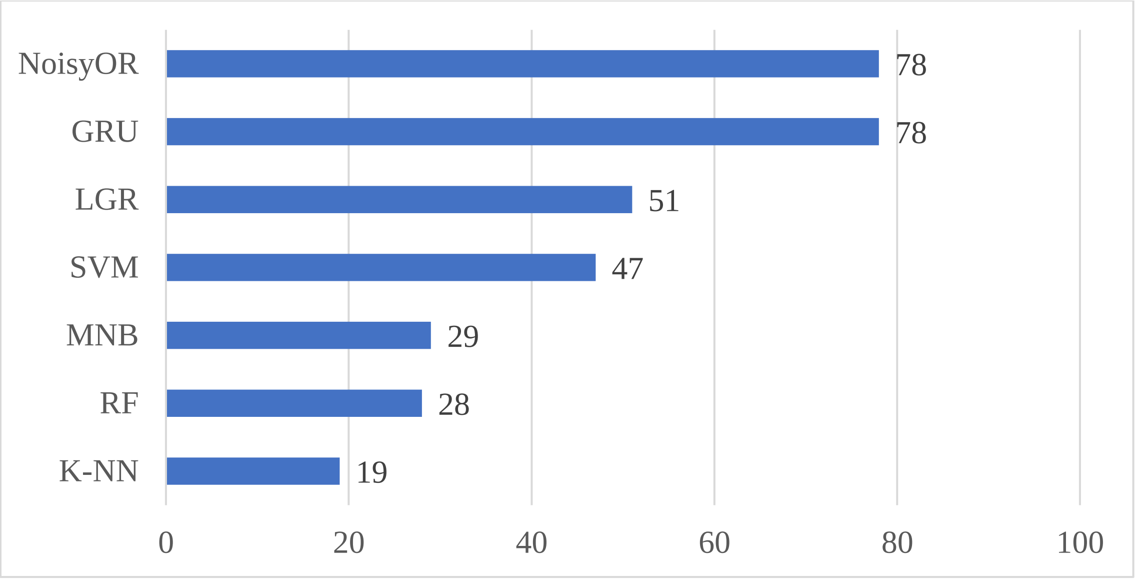

Figure 5-2 Missed WZ Crash Detection Accuracy (%)

It can be seen from Figure 5-2, Noisy-OR and GRU performed comparably, and each detected 78 WZ crashes, the highest number of missed WZ crashes among all. The performance is moderate for LGR and SVM but is not satisfactory for MNB and RF. K-NN classifier is the worst. MNB and RF provided more than 100 cases with a classification score of 1.0, most of which were NWZ crashes. It should be noted that doing well in the second evaluation is more difficult than doing well in the first because in the second evaluation the methods had to consistently not misclassify a narrative as WZ crash with a high score for a total of 82,215 narratives. Additional analysis is conducted for comparing unigram vs. unigram+bigram methods, identifying common cases from different models, and comparing the performance of Noisy-OR and GRU.

5.1.5.1 Comparing Unigram and Unigram+Bigram Methods of Noisy-OR

Section 5.1.4 explains that the unigram method may not be effective as expected for the Noisy-OR method. Positive unigrams with high cf values may have low probability values because the same unigrams also appear in NWZ crash narratives. The problem can be mitigated by adding some context to the Noisy-OR approach, such as in the form of bigrams. The ordered positive bigram list provides more contextual information related to WZ. This section provides empirical evidence of using the Noisy-OR method as a text classifier to identify missed WZ crashes from narratives. The section also explores the classification outcomes of unigram and unigram+bigram when compared with gold label, or manual reviewing.

The 100 narratives with the highest probability scores in each classifier were manually reviewed. The top 100 narratives of the unigram Noisy-OR classifier included 65 actual WZ crashes, while the top 100 narratives of the unigram+bigram Noisy-OR classifier included 78 actual WZ crashes. The unigram+bigram noisy-OR narratives that were correctly classified contained more contextual positive bigrams such as “construction zone”, “under construction”, “construction worker” and “lane closed” with high df values in the WZ training set.

A close review of 35 unigram Noisy-OR cases that were misclassified shows that they contain WZ-related positive unigrams such as “construction”, “barrels”, “attenuator”, “barricades”, “orange” and some noisy words such as “carrao”, “kampo”, “melloch”. These noisy unigrams have high df values in the WZ training set, indicating their popularity in the WZ crash narratives. On the contrary, the unigram+bigram Noisy-OR misclassified 22 cases from its top 100 narratives. A close review of these 22 cases reveals that the unigram portion of unigram+bigram Noisy-OR contains few positive unigrams but with high probability scores; the bigram portion contains a longer list of positive bigrams with moderate probability scores. Thus, the comparison reaffirms that unigram+bigram Noisy-OR tackled the noisy tokens more successfully than unigram Noisy-OR.

5.1.5.2 Analysis of Overlapping Cases for LGR, SVM, Noisy-OR and GRU

It is expected that in our second evaluation, many top scored narratives will be common among the models. However, we were surprised to see that there were only 333 (about 48% of the selected 700 narratives) overlapping narratives (narratives found in top 100 of one classifier were also found in top 100 of other classifiers). By analyzing overlapping narratives, we can gain insights into the classification performance of different classification methods on the same narratives. Therefore, we conducted a comparison study on the overlapping cases identified by LGR, Noisy-OR, SVM and GRU; the remaining models were not included in this analysis due to their poor performances.

According to Table 5-3, there are only 43 overlapping true WZ cases between Noisy-OR and GRU, which indicates that together, they found 70 different WZ cases (78-43=35 for each). Through the manual review, we found that these 70 cases contain reliable WZ keywords and are easy to classify, but when one method finds them among its top 100 narratives, the other does not. Another interesting observation is that out of the 47 true WZ cases of SVM, 45 overlap with LGR, and the model performance of SVM is very similar to LGR. There are 19 overlapping true WZ cases detected by all four classifiers (LGR, Noisy-OR, SVM and GRU) in which “construction zone”, “construction work”, and “construction lane” are the primary keywords (tokens).

Table 5-3 Overlapping True WZ among LGR, Noisy-OR, SVM and GRU

Classifier	Total	Overlapping Cases			
		LGR	Noisy-OR	SVM	GRU

Noisy-OR	78	45	78	41	43
GRU	78	20	43	20	78
LGR	51	51	45	45	20
SVM	49	45	41	49	20

By analyzing the 70 (i.e. 35+35) narratives as mentioned previously, we observed that the average length of narratives of Noisy-OR is longer than that of GRU. Figure 5-3 shows the distribution of narrative lengths for GRU and Noisy-OR. The GRU uses almost all the words in the narratives whereas Noisy-OR consider only positive words (i.e. unigrams and bigrams). A longer narrative may have many positive and non-positive words. As GRU consider all the positive and non-positive words, the overall classification score of the narrative may not be high. On the other hand, as Noisy-OR only considers positive words, the classification score will be increased. For example, if a long narrative has an equal number of positive and negative words and suppose GRU regards them equally, the classification score will be 0.5 for GRU, whereas it will be more than 0.5 for Noisy-OR. In this way, a longer narrative with or without many negative words is handled better by Noisy-OR. The GRU classified smaller narratives more accurately than Noisy-OR. In our dataset, we found that the probability score of the word "constructions" is 0.89 in Noisy-OR. If a narrative contains this positive word only with other non-positive words, the classification score of Noisy-OR will be 0.89 and it will not be included in the top 100 narratives because the classification score is above 0.99 for the top 100 narratives of Noisy-OR. Therefore, the narrative is not present in the top list of Noisy-OR. On the other hand, GRU gave high weight to some of the WZ related words. Therefore, the narrative with those words has higher GRU scores and consequently they are found in the top list of GRU. In short, Noisy-OR is sensitive to the number of positive words but not negative ones, whereas GRU is less sensitive to the number of positive words but can be affected by negative words.

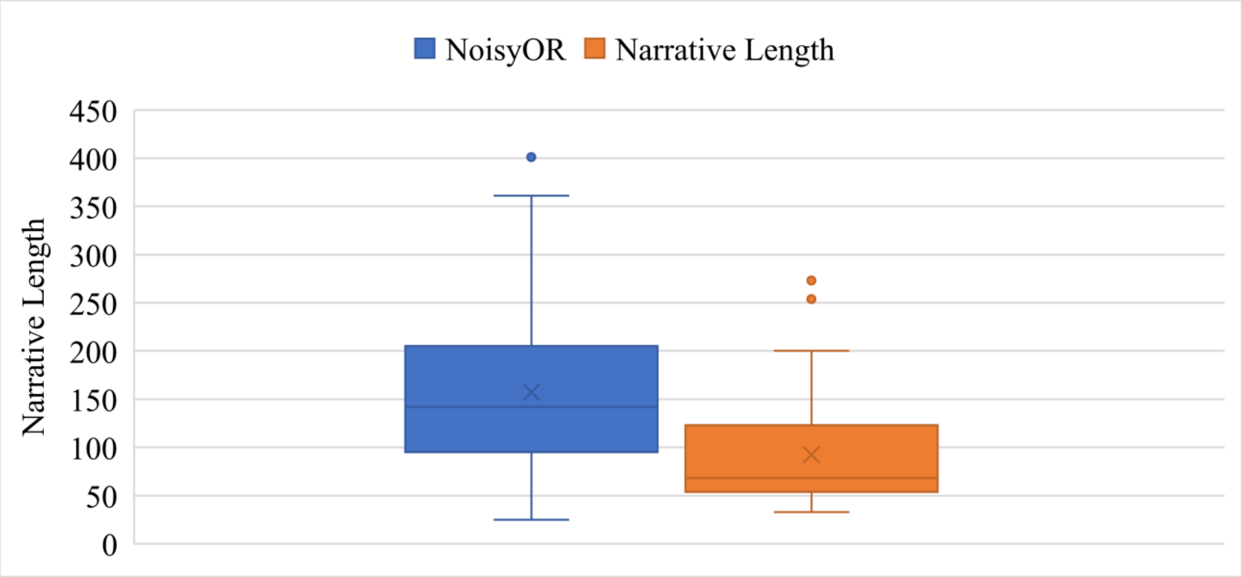

Figure 5-3 Distribution of Narrative Length

5.1.5.3 Comparison between GRU and Noisy-OR Results

Further investigation was done to the results of the top two performers: GRU and Noisy-OR. We found the top 100 cases of GRU result contained the word “construction” at least once in the naratives. We also found phrases in the narratives that lead to incorrect classification by GRU. These phrases can be categorized into two classes: mixed phrase and pesudo-WZ phrase. A mixed phrase refers to the combination of WZ related words and irrelevant words such as “johnson construction”, “construction at sarah's dance studio”, “construction on their new driveway”, and “construction building”. A pesudo-WZ phrase refers to the combination of WZ related words such as “construction barrels”, “construction equipment”, “construction barrier”, and “construction sign” but not in a work zone setting. Following are two example narratives that use pseudo-WZ phrases:

Narrative 1: *“Unit 1 was eastbound on i-94 in a snow storm lost control went into the median struck some* ***construction barrels*** *and ended up on the westbound side of i-94”*. In this narratives, only the presence of construciton barrels does not warrant that there was a construction zone in the travel direction.

Narrative 2: *“Unit one was traveling westbound (north) on us 14 just south of sth 138. Unit one struck a ladder which was present in the middle of the roadway. Unit one continued to travel westbound, where he observed a white-colored pickup truck (with* ***construction equipment*** *in the bed) making a u-turn at the turn around on us 14 and netherwood st. Unit one followed the pickup truck after turning around, and later confronted the driver of the pickup truck over the ladder. The driver of the pickup truck denied the ladder was his. I was able to speak with the driver of the pickup truck a few days later, and he denied the ladder was his. He advised he works for a roofing business and owns ladders significantly larger than the one that was present* *in the roadway".* In this narrative, it is clear that the construction equipment was loaded in a pickup but the accident has nothing to do with a work zone.

Among the 78 correctly classified narratives of Noisy-OR, 75 narratives contain the word "construction". Among the 22 misclassified narratives of Noisy-OR, the word "construction" appeared in the 9 narratives, and the words "barrels", "barrier", and "orange" appeared several times. The pesudo-WZ phrase "construction barrels" appeared in 7 narratives. Since Noisy-OR is a keyword-based classifier and does not use contextual information for classification, it fails to correctly classify these narratives. Furthermore, we found that the words "lane", "closed", "attenuator", "orange", and "barrel" appeared several times in the remaining misclassified narratives.

The above analysis shows that both GRU and Noisy-OR perform well, but their classification mechanisms are different. The reasons of misclassifications are very similar for some of the cases (e.g., presence of pesudo-WZ phrase). However, it is difficult to select the best classifier based on manual review of the top 100 results of GRU and Noisy-OR. That is why, we expanded the sample size from top 100 to top 200. Figure 5-4 shows the detection rate of missed WZ crashes in an interval of 50 data points with a maximum of 200 narratives. We found that GRU detected 146 missed WZ crashes whereas Noisy-OR detected 137 from their top 200 narratives. The detection rate of GRU fluctuates with the decrease of the classification score, but for Noisy-OR, it decreases with the decrease of the classification score.

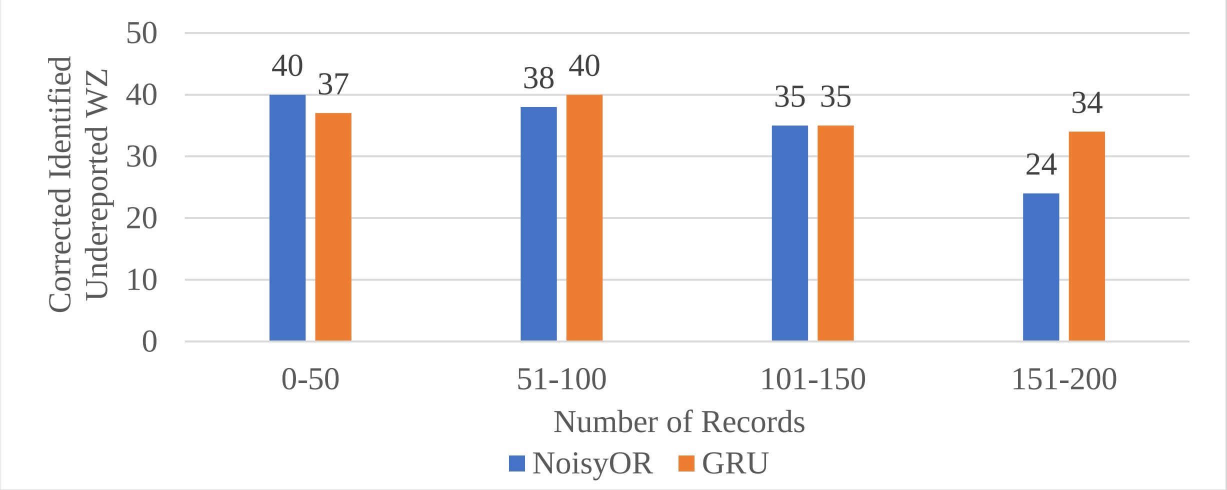

Figure 5-4 Accuracy of Noisy-OR and GRU

5.1.6 Extended Analysis of Unigram+Bigram of Noisy-OR

Further analysis was performed to quantify the classification accuracy rate against the case rank of the unigram+bigram method. Starting from the highest-ranked cases, the number of correctly identified WZ crashes is counted over the 50-case intervals, as shown in Figure 5-5.

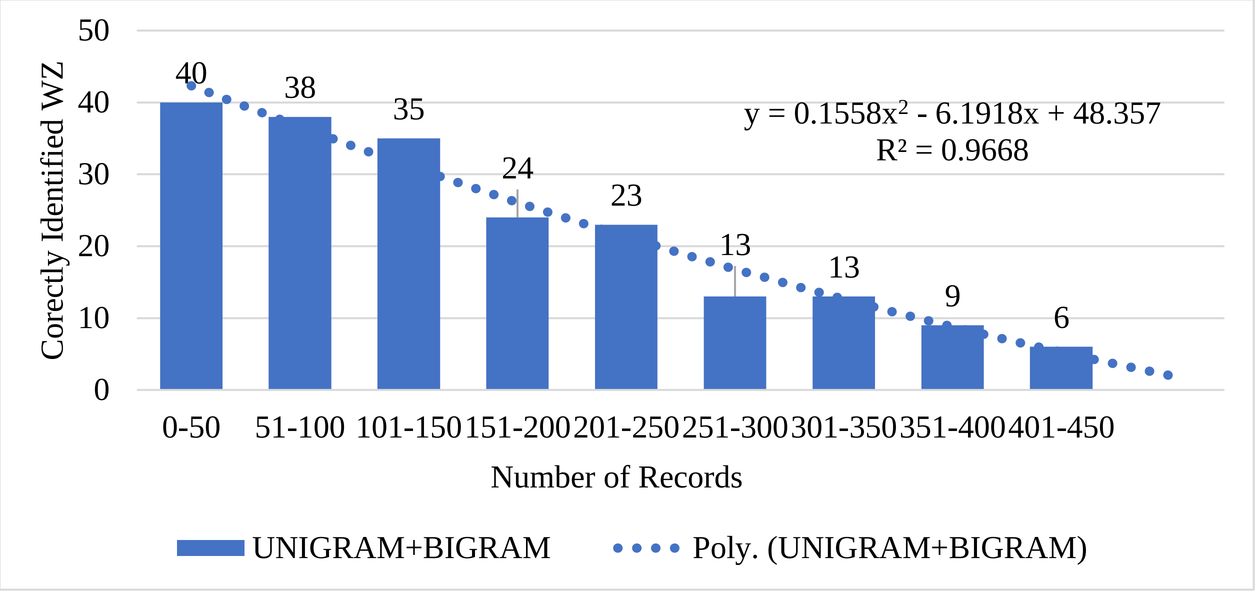

Figure 5-5 Accuracy of (Unigram+Bigram) Noisy-OR

From Figure 5-5, two observations can be made based on the 450 cases reviewed: a) more than 50% of cases correctly classified till the fifth interval (201-250), and b) the model performance degrades rapidly from 80% in the first interval [0-50] to 12% in the last interval [401-450]. The fitted quadratic Equation has a R^2 value of 0.9668, suggesting a strong and consistent trend for the descending accuracy rate. The findings are good news for an agency who wants to estimate the effort of a manual review for missed WZ crashes.

The probabilistic distribution of narrative length was plotted for WZ and NWZ crashes, respectively, in Figure 5-6. The distribution was inspired by a study that shows that narratives not designated by officers as speed-related crashes have a longer length on average than non-speed related crashes (Fitzpatrick et al. 2017). Figure 5-6 shows that the narrative length of actual NWZ crashes is approximately normally distributed, while missed WZ crashes are slightly skewed toward the left. The two distributions are statistically different at a 5% level of significance (two sample t-test, p=<0.0001).

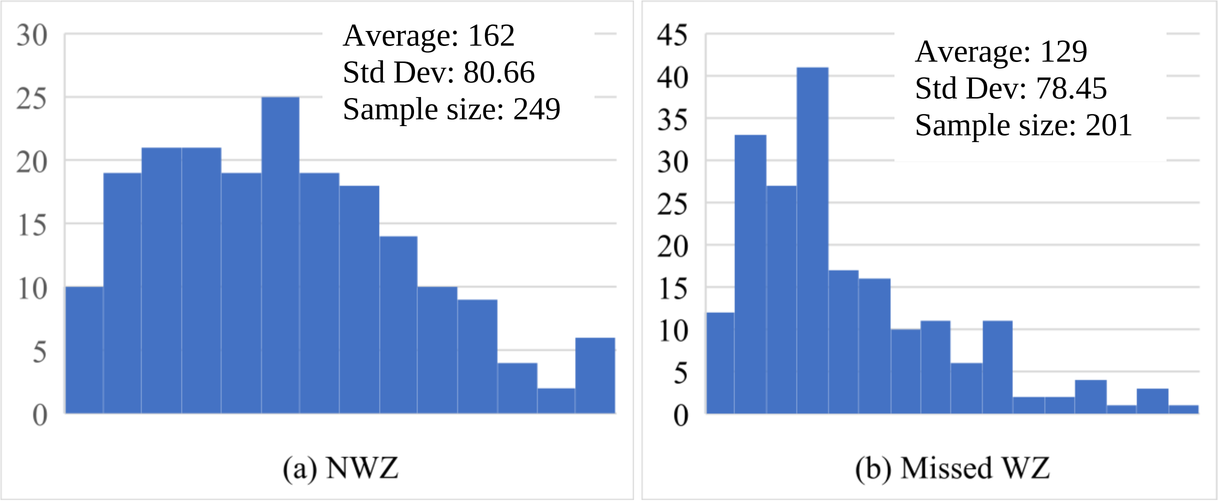

Figure 5-6 Histogram of Narrative Length for a) NWZ and b) Missed WZ

Moreover, the average narrative length of reported WZ crashes is 104, and Std. is 68 (sample size:1989), which is a statistically significant difference between NWZ (two sample t-test, p=<0.001) and missed WZ (two sample t-test, p=<0.001). Though it is expected that long narratives would have more positive tokens than short narratives, no correlations are observed between the length of narratives and the number of positive tokens for reported WZ and NWZ and missed WZ. In other words, there is not enough evidence to claim that long narratives tend to classify crashes more accurately than short narratives.

5.1.7 Summary of Observations

We obseve that there are two main challenges in identifying missed WZ crashes from narratives. The first challenge is due to the nature of WZ crash narratives. A crash narrative may be very long with several parts irrelevant to WZ, but if it mentions a word or a few words, such as “construction area”, at just one place then that would make it a WZ crash narrative. In additon, there are not many words or phrases which are indicative of WZ. This is typically not how most classes are in text classification tasks. For example, for a classification task to classify a news article as belonging to politics or not, most parts of the article will indicate that it belongs to politics and there will be plenty of words that will be indicative of that class. Many popular text classification statistical methods, such as SVM, LGR and MNB, work well with the latter types of tasks, because they tend take into account a large number of features to make their classification decisions. On the other hand, Noisy-OR can narrow down to only a few indicative words and only looks for their presence, and because it is a probabilistic “Or”, presence of any good indicative word is sufficient for it to classify a narrative as WZ. This is one reason why Noisy-OR worked well on our task. GRU’s learning mechanism is complex and not easily interpretable, but it appears from the results that it also learned to base its decision on the presence of a few indicative words.

One more reason the nature of WZ crash narratives is different from typical text classification classes is that they can very well contain what can be in NWZ crash narratives. This is because what happens in an NWZ crash can also happen in a WZ area thus making it a WZ crash. In contrast, for example, a non-political news article will be always very different from a politcal news article. In other words, there are really no negative words that indicate that a narrative is not WZ. However, methods such as LGR and SVM heavily use negative features (e.g. Figure *5-1*) which possibly confuse the methods on this task. On the other hand, Noisy-OR strictly uses only positive indicators and hence is not affected. It appears from the earlier discussion related to the lengths of narratives that GRU is affected to a some extent by the negative features.

The second challenge is due to the way we automatically created the training dataset that led to large noise as was pointed out earlier. An estimated 70% of narratives flagged as WZ do not contain anything that indicates WZ crash. This adversely affects most of the methods because they are not designed to handle so much noise. In contrast, this is unlikely to affect Noisy-OR's top unigrams and bigrams as long as there are sufficient true WZ narratives flagged as WZ. From the results, it appears that GRU was not much affected by this noise. There is also noise because many narratives flagged as NWZ are, in fact, the missed WZ crashes. Although this can potentially affect all the methods, given that a small percentage of all crashes are WZ, the extent of this noise is small. Although the above observations are specific to the task of identifying missed WZ crashes, it is likely that they will be true for the tasks of identifying other missed causes of crashes. To summaize the results, among the seven classifers tested, MNB, RF and K-NN provide poor classifiction performance with our dataset. Although the AUC score, and the coefficients of WZ-related words of LGR and SVM seem promising, the performance of LGR and SVM in detecting missed WZ crashes is not satisfactory for the reasons mentioned earlier. GRU and Noisy-OR are the two best performers, and their results of recovering missed WZ crashes from the reported NWZ crash narratives are comparable. Based on manual verification of the first 200 narratives of each model, GRU detected 146 WZ crashes, 9 more WZ crashes than Noisy-OR. Noisy-OR can handle longer noisy narratives better than GRU. On the other hand, compared to Noisy-OR, GRU can handle shorter narratives better. The word probability of a positive word in Noisy-OR is prepared in such a way that if an important positive keyword is very frequent in the NWZ narratives, the word probability score is decreased. Therefore, although some short narratives of missed WZ crashes have obvious indication of WZ crashes, Noisy-OR may not be able to generate higher classification scores for the narratives. On the other hand, GRU does not emphasize much on the number of occurrenes of keywords in the narratives. Instead, it uses the context of the words through its sequence processing mechanism.Therefore, it is able to correctly classify short narratives. However, GRU is not able to generate high classification scores for the longer narratives. This indicates that GRU cannot handle narratives that have a few positive words with many negative words. But GRU has the advantage that it employs word embeddings which enables it to treat semantically similar words similarly in its model (for example, it will treat "barricade" and "roadblock" similarly). But Nosiy-OR treats every word distinctly whether they are semantically similar or not. On the other hand, Noisy-OR is simple, computationally fast and interpretable. Whereas, GRU algorithm is very complex in nature and requires fine-tuning several hyperparameters. It also requires significant amount of time to train the model. Therefore, there is a trade-off in choosing the best model between Noisy-OR and GRU.

Considering model complexity and computational power, the unigram+bigram noisy-OR method is an effective and efficient method for classifying text and recovering missed WZ crashes for real-time application. According to Figure 5-5, a review of the top 450 cases of the unigram+bigram noisy-OR identified 201 WZ crashes as missed, which is more than 8% of reported WZ crashes from 01/01/2019 to 10/31/2019. Moreover, the decreasing trend of finding missed WZ crashes suggests the chance may be 12% or lower after the first 450. Additionally, 450 crashes is a tiny fraction of the pool of potentially missed WZ crashes (i.e., 125,509 NWZ crashes in 2019), which is very helpful to an agency that wants to prioritize and estimate the level of effort of a manual review.

5.1.8 Spatial-Temporal Analysis of Missed WZ Crashes

Further analysis was conducted on the crash time and location for a better understanding of the circumstances under which a WZ crash is missed. Figure 5-7 shows the distribution of reported WZ and missed WZ confirmed in this study by time of day, day of week, and month of year.

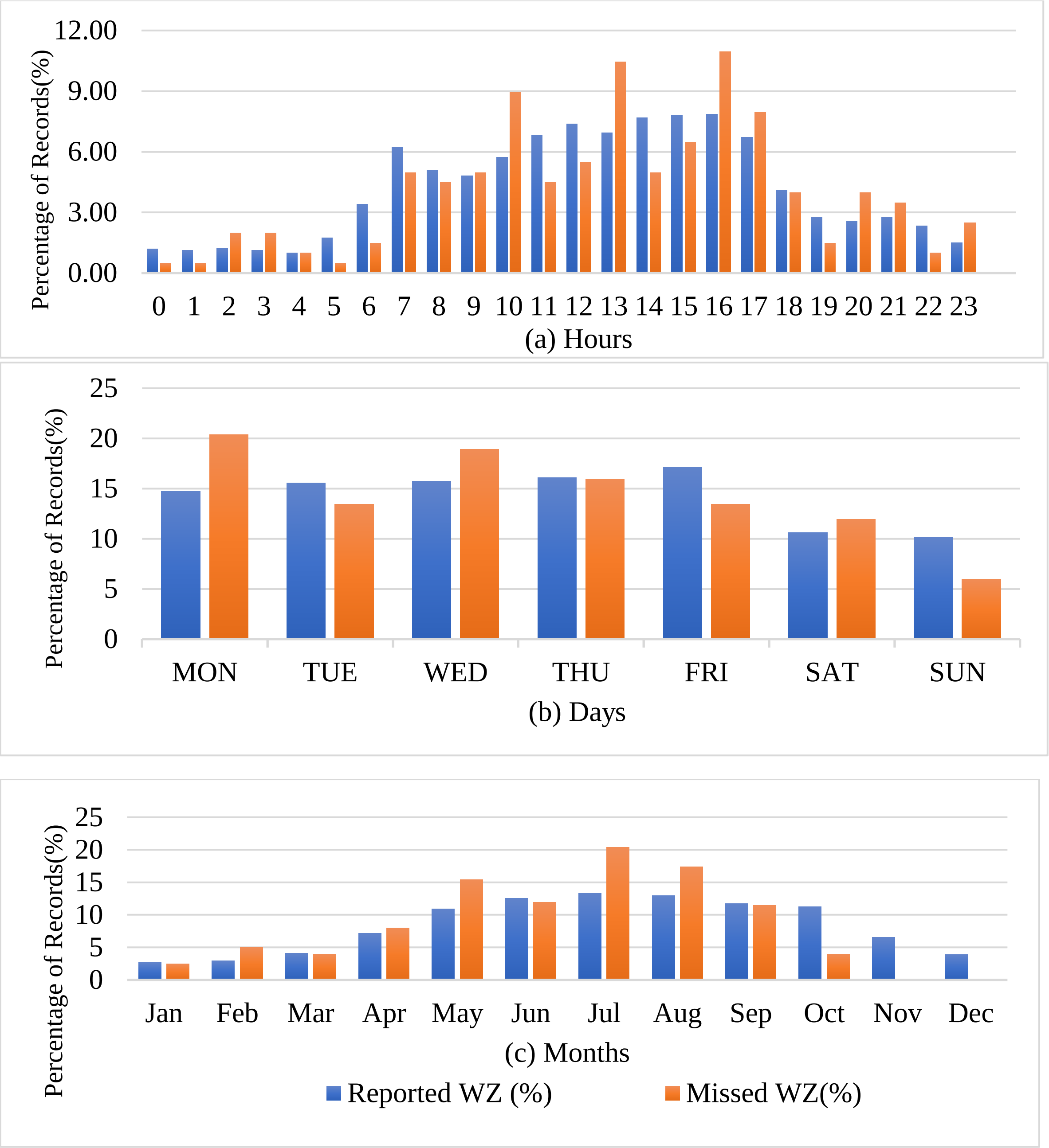

Figure 5-7 WZ Crash Analysis by a) Hour b) Day and c) Month

In 2017 to 2019, 70.96% of all reported WZ crashes and in 2019, 73.13% of the missed WZ crashes identified in this study occurred during daylight hours from 8 a.m. to 6 p.m., as shown in Figure 5-7 (a). Among daytime WZ crashes, a high percentage of missed cases occurred in the afternoon when traffic is busiest, from 4 p.m. to 5 p.m. It is plausible that crashes are missed when traffic is high or when construction activities are intense. The day of week distribution suggests that the WZ crashes are probably missed throughout the week, especially on Monday and Saturday, as shown in Figure 5-7 (b). Figure 5-7 (c) also displays the monthly distribution of reported WZ crashes versus missed WZ crashes, showing that a high percentage of missed cases are observed in the summertime, especially in July and August when construction activities are extensive and intensive.

Figure 5-8 shows the distribution of missed WZ crashes compared to reported WZ crashes by highway class. The evidence shows that most missed WZ crashes occurred in urban areas, including urban city streets (43.11%), urban state highways (16.89%) and urban interstate highways (15.33%). The interstate highway system, both urban and rural, has the best performance in terms of a low ratio of missed crashes to reported crashes. The next best performance is from state highways, where the ratio is close to 1. City streets have the highest ratio of missed crashes to reported crashes, particularly urban city streets which have only 20% of the total reported WZ crashes but make up 43% of missed WZ crashes identified in this study. Cheng et al. stated that construction work zones are usually assumed to be long term works, but maintenance or utility works are usually short term and temporarily, which may not be known to driver in advance (Cheng et al. 2012). Since many crashes on urban streets involve utility work zones, it is plausible that police may not consider those as construction zone related.

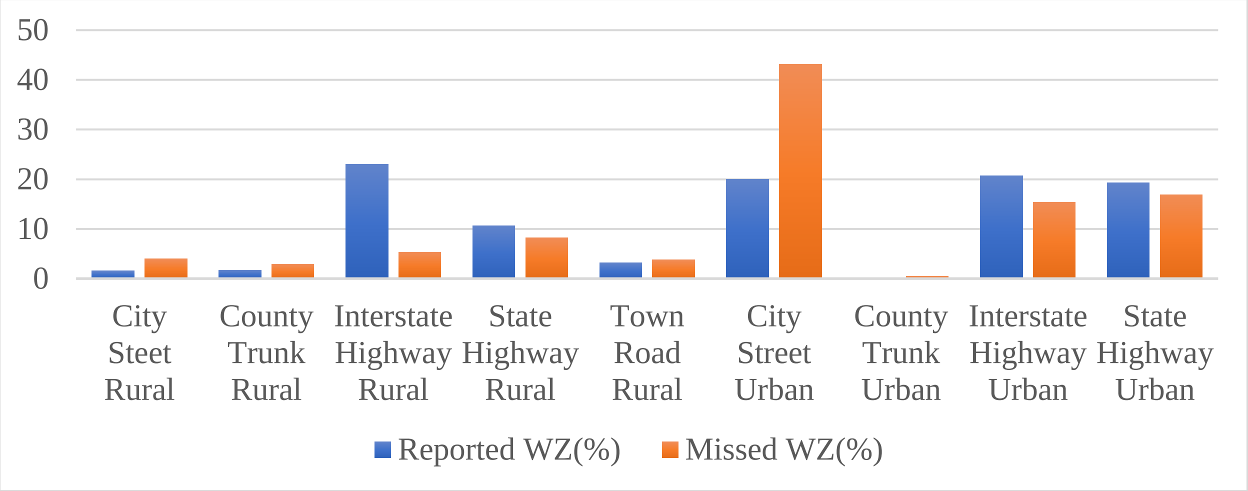

Figure 5-8 WZ Crash Analysis by Highway Class

Comparisons were conducted for other structured data fields, including weather conditions, pavement conditions, light conditions, and injury severity. The results show similar distributions between all reported WZ crashes and missed WZ crashes, mainly due to the lack of variety since most WZ crashes, reported or missed, occur during clear or cloudy weather, on dry pavement, in the daytime, and involve less severe injuries.

An analysis of missed cases suggests the 73.13% of the missed WZ crashes identified in the study occurred from 8 a.m. to 6 p.m. with a high percentage in the afternoon from 4 p.m. to 5 p.m. A high percentage of WZ crashes that are misclassified are observed in July and August when the construction activities are extensive and intensive. 43% of the missed WZ crashes identified in this study occurred on urban city streets.

5.2 Distracted and Inattentive Driving Crashes

There are many reasons behind car crashes, including driver distraction and inattentiveness. In Wisconsin, USA alone, from 2017 to 2019 the percentage of crashes due to distraction and inattentiveness increased from 8.28% to 12.41%, according to the crash statistics from Wisconsin reportable crashes. When a driver fails to pay sufficient attention to perform basic tasks for safe driving, the driver is called inattentive, and the driving is called inattentive driving. While there is only action or activity behind inattentive driving, distracted driving involves both action and a source of distraction.

The definition and categorization of driver distraction and driver inattention can be referenced from existing sources and research. (R. Dewar and P. Olson 2007) stated that “the essential distinction between inattention and distraction is that inattention is internal to the driver and non-compelling, whereas distraction is external to the driver and compelling”. (Regan, Lee, and Young 2008) stated that *the absence (in the case of driver inattention) of a competing activity* is the key factor in differentiating driver distraction from driver inattention. (Hoel, Jaffard, and Van Elslande 2010) distinguished driver inattention from driver distraction according to the nature of the competing activity. For driver distraction this is any external non-driving-related activity and for inattention, this activity is preoccupation in internalized thought. (Regan, Hallett, and Gordon 2011) defined distracted and inattentive driving, found relation between them, and made a taxonomy for them. They defined driver inattention as “insufficient or no attention to activities critical for safe driving” and categorized driver distraction as another form of driver inattention. According to the National Highway Traffic Safety Administration (NATSA), distracted driving is defined as “Distracted driving is any activity that diverts attention from driving, including talking or texting on your phone, eating and drinking, talking to people in your vehicle, fiddling with the stereo, entertainment or navigation system — anything that takes your attention away from the task of safe driving” (NHTSA 2021). In summary, driver distraction involves a triggering event, a competing activity; where the competing activity is externally generated and may lead to attention shift.

Though distracted driving is considered as a specific type of inattentive driving (NHTSA 2010), the growing crash reports due to distraction lead us to consider them separately. Distracted driving involves some internal (i.e., inside the vehicle, for e.g., phone, radio, gps, etc.) or external sources, where inattentive driving does not involve any sources. It is important to differentiate the two types of crashes because knowing the source of distraction can help us take appropriate intervention. Furthermore, specific safety treatments for distracted or inattentive driving related crashes can be implemented for improved effectiveness.

5.2.1 Data Collection

In this study, crash reports were acquired from the Wisconsin Department of Transportation (WisDOT) through the WisTransPortal data hub, including all the crash narratives. We collected data during a transitional period of the database and observed several changes in the data elements, which are described below.

In the dataset before 2019, the field "DISFLAG" referred to all distracted and inattentive driving crashes (DOI). Therefore, the narratives that marked under DISFLAG flag could not describe which one was distracted and which one was inattentive narrative. In 2018, new data elements were added to the database to help separate distracted driving from inattentive driving. The implementation was rolled out gradually as law enforcement agencies upgraded their computer systems. Therefore, the database was not complete during that transitional time. In 2020, we collected the data for the year 2017 to 2019 and found that there were 32,050 DOI crashes. In order to prepare training dataset for DD and ID cases, we had to manually annotate some of the DOI cases as DD, ID, DD+ID and ND cases. However, manually separating DD and ID from the huge DOI data set is not an effective method. It is expected that data collected in a later time after the crash data improvement project such as 2019 to 2021 data are far better in distinguishing DD and ID from DOI with specific elements. That is why, after having a complete dataset for the years 2019 to 2020 and a partially complete dataset for the year of 2021 (till June 16), the distracted and inattentive crashes are populated based on the following query.

- *Distracted Crash* when {DISTACT [1,2] are not “Not Distracted” and not blank} & {DISTSRC [1,2] is not “Not Applicable”) (Not Distracted) and not blank} then it is Distracted crash, otherwise not
- *Inattentive Crash* when {ID} in DRVRPC [1,2] [A, B, C, D] then it is Inattentive crash, otherwise not

The field “DISTACT” provides the actions of drivers such as talking, listening, manipulating or other actions. The field “DISTSRC” provides the source of distraction such as hands-free mobile phone, hand-held mobile phone, vehicle-integrated device, or other source of distractions. The “DRVRPC” provides both inattentive driving and distracted driving parameters.

Table *5-4* presents the overall crash statistics after populating DD and ID crashes from 2019 to 2021. We found 51,405 DD cases and 17,791 ID cases in 2019-2020 dataset. From the experience of work-zone crash classification (section 5.1Work Zone Crashes), it can be said that both datasets represent a good amount of data for using as training dataset in Noisy-OR. Therefore, we developed individual classifiers for both DD and ID cases.

Table 5-4 Crash Statistics

Year	Type	Distracted (%)	Inattentive (%)
2019	Reported	27,135 (21.16%)	9,994 (7.79%)
	Not Reported	101,074 (78.84%)	118,515 (92.21%)
	Total	128,209 (100%)	128,209 (100%)
2020	Reported	24, 270 (23.9%)	7,797 (7.68%)
	Not Reported	77,239 (76.1%)	93,712 (92.32%)
	Total	101,509 (100%)	101,509 (100%)
2021 (partial dataset)	Reported	10,905 (23.72%)	3,612 (7.86%)
	Not Reported	35,077 (76.28%)	42,370 (92.14%)
	Total	45,982 (100%)	45,982 (100%)

For the DOI classifier, the DOI cases can be prepared from the DD and ID narratives in two ways: (1) either distracted or inattentive cases + both distracted and inattentive cases (2) both distracted and inattentive cases. Since our dataset is very noisy like work zone cases, the former method (either distracted or inattentive cases) will add noisy narratives to the training set from DD and ID cases. On the other hand, the other method (both distracted and inattentive cases) will reduce the noisy narratives in the dataset because it is unlikely for a narrative without any DD or ID related words to be reported as a both inattentive and distracted case. Both methods were used; and the one that provided higher accuracy to classify DOI cases was chosen.

We started with the data from 2019-2020 as training data for all the classifiers. However, the classifiers for the cascade classifiers (models) cannot be tuned because the training data is very noisy (the narratives do not contain any DD, ID or DOI related words). Alternatively, we randomly selected 300 narratives for DOI, 500 narratives for DD and 500 narratives for ID from 2018-2019 dataset and manually annotated them to find the optimal threshold values of the classifiers (the details are described later). The use of manually annotated data ensures that these crashes are properly classified.

We used 2020 dataset as the preliminary test data to investigate how well the classifiers are trained. We took the top 100 results of each classifier, and manually investigated them to get deeper insight about the classifiers. The 2021 dataset was used as final test dataset for all the classifiers to investigate how well the classifiers performed in a new dataset.

After obtaining the DOI, DD and ID classifier, we prepared cascades classifiers (models) by combining them, as described in section 4.7.3. To prepare the test data for the models, the distracted (DD) and inattentive (ID) cases was prepared by manually reviewing 300 narratives (reported and unreported) that were randomly taken from 2018-2020 dataset. We found 93 distracted (DD) and 87 inattentive (ID) and 120 Neither (NDOI) cases. To make a balance dataset, we selected 90 Neither (NDOI) cases randomly from the 120 NDOI cases.

The following three sample crash narratives were chosen randomly from the dataset to illustrate the structure of crash narratives; the first one is distracted driving (DD), the second one is inattentive driving (ID), and the third one is NDOI.

- DD crash narrative example: *Both units were traveling w/b on w Ryan Rd and approaching 22nd St when unit #2 stopped at the red light. The driver of unit #1 stated that she was on her way to a job interview and looked at her GPS, when she looked back up she thought it was a green light, but it was actually a red light. Unit #1 was not able to stop in time and struck unit #2 in the rear bumper, causing minor damage to unit #2 and disabling damage to unit #1. B&b towed unit #1.*
- ID crash narrative example: *Unit one was driving eastbound on interstate 94 in the left lane. The vehicle deviated from its lane crossing the fog line and striking traffic cones on the median shoulder. Upon striking the cones the vehicle swerved back onto the road but over corrected. It caused the vehicle to spin to the right going into the right lane. The vehicle the left the roadway onto the right shoulder sideways and then turned completely where the rear of the car struck a dividing fence. The dividing fence is what separates the interstate from a corn field. When I asked the driver what happened he stated he was tired and dozed off. When he fell asleep for that split second, he had swerved to the left going off the road.*
- NDOI crash narrative example: *On 01/05/2019 at 2018 hours, I responded to report of a two vehicle hit and run crash. I met the reporting party at her residence. The RP, driver of unit 1, was traveling sb on 17th av, and made a left turn onto Sherman St. Unit 1 was struck by a vehicle traveling nb on 17th av. The striking vehicle did not stop. Unit 1 drove the vehicle home before calling. She believed the striking vehicle was a dark colored sedan, but had no further information. Unit 1 believed she had a green light but did not know if she had a green arrow. Damage to right rear of unit 1. Observed some small plastic pieces scattered around intersection, but no identifying features for the striking vehicle. Photographs attached via axon capture.*

5.2.2 Result and Analysis

In this section, the unigrams (U), bigrams (B), and trigrams (T) with the highest probability scores and thus strong indicative of a DOI, a DD, or an ID narrative are presented. Next, the U, U+B, and U+B+T approaches are compared, and the best one is used in the hierarchical and priority models. Finally, the results of models with DOI, DD, and ID classifiers are calculated and compared.

5.2.3 Unigram, Bigram, and Trigram Probability Analysis

Table 5-5 **Table *5-5***to Table 5-7 show the top 25 unigrams, bigrams, and trigrams with their corresponding probability, positive count, and negative count for DOI, DD, and ID. In these tables, DOI/NDOI means word count in DOI vs. word count in non-DOI cases (NDOI); same for DD/NDD and ID/NID. All calculations were performed by the Equation of weighted count probability (Equation 3). It is clear that the method successfully obtained the most relevant unigrams, bigrams, and trigrams for each narrative type.

Table 5-5 lists the most important words that are prepared from distracted and inattentive narratives and that have the highest probability scores. For example, the words “inattentive” and “distracted” has a probability of 0.99 and 0.97, respectively. With the Equation of simple count probability, however, the two words have a lower probability of 0.81 and 0.69, respectively (calculated separately, not shown in Table 5-5). Therefore, the Equation of weighted count probability works well to extract the most critical unigrams/bigrams/trigrams and gives them high weights (probability). Some common phrases from the DOI narratives are (based on our manual review): not looking on the road, not paying attention, inattentive driving, operating phone/radio/GPS, reaching for drink/dropped phone or object, adjusting radio/visor, etc. Overall, Table 5-5 shows a good reflection of these phrases in all unigrams, bigrams, and trigrams. With these high probability scores of essential words, a test narrative will be given a high Noisy-OR score when these words are present and thus classifying DOI cases from NDOI cases.

Table 5-5 Top 25 Uni-, Bi, and Tri-grams for the DOI Classifier

Unigram	Pro	DOI/NDOI	Bigram	Pro	DOI/NDOI	Trigram	Pro	DOI/NDOI
inattentive	0.9913	1632/391	inattentive driving	0.9907	1491/351	for inattentive driving	0.9888	1266/298
distracted	0.9727	931/409	for inattentive	0.9890	1290/303	cited for inattentive	0.9726	662/137
paying	0.9501	601/323	looked down	0.9813	845/183	inattentive driving unit	0.9314	383/78
cell	0.9266	443/256	was distracted	0.9610	544/128	not paying attention	0.9311	412/167
looking	0.9222	1557/1550	looked up	0.9581	599/250	citation for inattentive	0.9231	360/86
gps	0.8555	256/150	distracted by	0.9513	480/124	was looking at	0.9138	373/194
phone	0.8516	1557/2261	down at	0.9496	527/228	when he looked	0.9109	408/278
radio	0.8263	355/468	paying attention	0.9472	561/295	he was looking	0.9085	379/235
reached	0.8244	267/282	not paying	0.9377	439/173	was distracted by	0.9081	319/60
looked	0.7987	2878/5072	looking at	0.9375	590/398	looked down at	0.9049	310/41
asleep	0.7888	473/800	was looking	0.9370	918/757	was not paying	0.8895	299/130

eyes	0.7706	318/525	cell phone	0.9279	432/229	he looked down	0.8826	275/71
dropped	0.7649	195/222	his phone	0.9274	413/196	when she looked	0.8752	285/159
reaching	0.7362	159/154	looked away	0.9271	376/103	she was looking	0.8501	253/162
grab	0.7345	142/48	phone and	0.9160	391/220	looked down to	0.8477	227/40
floor	0.7075	149/181	looked back	0.9098	429/319	driving unit 1	0.8441	298/301
notice	0.7006	254/512	looking down	0.9056	317/79	looked away from	0.8413	220/35
bottle	0.6734	114/84	her phone	0.8994	318/133	inattentive driving and	0.8378	218/53
texting	0.6708	108/27	driving unit	0.8958	470/443	he looked up	0.8378	224/96
talking	0.6599	144/256	at his	0.8823	419/413	was looking down	0.8127	195/50
watching	0.6507	136/243	at her	0.8696	349/330	down at his	0.8093	191/34
attention	0.6435	968/2522	he looked	0.8382	951/1431	she looked down	0.8054	188/35
inattentively	0.6239	86/22	eyes off	0.8330	213/50	paying attention and	0.7942	189/112
cigarette	0.6145	82/30	attention to	0.8320	226/133	looking down at	0.7911	177/32
coffee	0.6141	88/95	she looked	0.8277	687/1052	she looked up	0.7896	178/59

Table 5-6 shows top unigrams, bigrams, and trigrams related to the DD classifier. From the unigram column, we can see that, except for some unigrams like inattentive, asleep, etc., all the words are a strong indicator of distracted driving like distracted, GPS, cell, radio, phone, reached, talking, dropped, grab, reaching, floor, bottle, etc. Though "distracted" has the highest probability, the word "inattentive" has the second-highest probability given that police officers do not differentiate between distracted driving and inattentive driving on their written narratives. From the manual review, we frequently noticed that even when a narrative presents a distracted driving case, the last line of the narrative is something like "unit # is cited for inattentive driving.". The frequent appearance of "inattentive" makes it very difficult to train a stable and effective DD classifier. The same analysis is true for the bigrams and trigrams of Table 5-5.

Table 5-6 Top 25 Uni-, Bi, and Tri-grams for the DD Classifier

Unigram	Pro	DD/NDD	Bigram	Pro	DD/NDD	Trigram	Pro	DD/NDD
distracted	0.9659	1906/523	looked down	0.9677	1450/266	for inattentive driving	0.9542	1912/662
inattentive	0.9560	2495/907	inattentive driving	0.9583	2284/786	cited for inattentive	0.9222	987/349
paying	0.9107	1071/479	for inattentive	0.9546	1953/678	was distracted by	0.8920	628/94
looking	0.8898	3025/1943	was distracted	0.9511	1080/197	not paying attention	0.8830	708/275
cell	0.8875	756/309	distracted by	0.9410	953/174	was looking at	0.8772	673/258
phone	0.8453	3247/2559	looked up	0.9329	1061/340	when he looked	0.8704	741/363
gps	0.8448	529/181	down at	0.9136	861/286	inattentive driving unit	0.8642	586/192

reached	0.8132	556/330	looking at	0.9102	1089/495	citation for inattentive	0.8525	548/181
radio	0.8116	713/522	paying attention	0.9075	1004/443	he was looking	0.8505	646/328
looked	0.7990	6103/5689	looked away	0.9029	696/147	looked down at	0.8497	498/73
eyes	0.7851	726/609	his phone	0.9017	767/252	was not paying	0.8352	529/220
talking	0.7500	413/280	was looking	0.8982	1725/1006	when she looked	0.8284	507/208
dropped	0.7391	386/259	not paying	0.8902	748/288	he looked down	0.8211	446/104
grab	0.7193	289/66	cell phone	0.8876	731/281	she was looking	0.8069	472/222
attention	0.7122	2371/2788	phone and	0.8858	711/266	he looked up	0.7842	393/135
reaching	0.7113	313/179	looked back	0.8827	822/396	driving unit 1	0.7813	558/417
asleep	0.7053	1057/1231	her phone	0.8565	553/171	looked down to	0.7812	368/57
notice	0.6960	547/594	looking down	0.8558	530/126	looked away from	0.7763	361/58
floor	0.6876	296/202	at his	0.8420	769/490	she looked down	0.7521	327/52
ejected	0.6813	366/350	at her	0.8307	639/383	paying attention and	0.7462	352/169
owi	0.6646	1382/1804	she looked	0.8234	1433/1178	inattentive driving and	0.7460	334/119
alcohol	0.6638	507/603	driving unit	0.8222	821/605	was looking down	0.7443	323/85
bottle	0.6547	228/104	he looked	0.8189	1920/1640	he looked back	0.7300	331/168
seat	0.6512	1927/2611	eyes off	0.7818	375/87	she looked up	0.7245	299/86
def	0.6452	724/959	attention to	0.7768	400/180	down at his	0.7229	292/58

Table 5-7 shows top unigrams, bigrams, and trigrams related to the ID classifier. From the unigrams column of Table 5-7, we see the reflection of the previous statement. The same observation is valid for the bigrams and trigrams column of Table 5-7. One important observation is that the highest probability for the unigram and bigram “inattentive” is 0.9642 and 0.9623, respectively; which is higher than that in Table 5-6. This indicates that the DD classifier is a less effective classifier to separate DD cases from ID cases with the presence of ID classifier.

Table 5-7 Top 25 Uni-, Bi, and Tri-grams for the ID Classifier

Unigram	Pro	ID/NID	Bigram	Pro	ID/NID	Trigram	Pro	ID/NID
inattentive	0.9642	2148/1254	for inattentive	0.9623	1665/966	for inattentive driving	0.9622	1630/944
paying	0.8816	757/793	inattentive driving	0.9623	1926/1144	cited for inattentive	0.9531	874/462
distracted	0.8435	1045/1384	looked down	0.9180	928/788	inattentive driving unit	0.9176	498/254
cell	0.8423	496/569	not paying	0.8894	554/482	citation for inattentive	0.9003	455/274
looking	0.8070	1950/3019	down at	0.8811	585/562	not paying attention	0.8819	520/463
asleep	0.8022	904/1384	paying attention	0.8798	709/738	looked down at	0.8558	342/229

gps	0.7400	287/423	looked up	0.8797	689/712	he was looking	0.8504	472/502
reached	0.7020	315/571	was distracted	0.8660	613/664	was not paying	0.8481	389/360
phone	0.6902	1855/3951	looking down	0.8529	364/292	was looking at	0.8340	437/494
radio	0.6860	409/826	distracted by	0.8503	530/597	inattentive driving and	0.8286	284/169
dropped	0.6755	232/413	cell phone	0.8486	484/528	he looked down	0.8286	308/242
grab	0.6699	160/195	driving unit	0.8470	634/770	when he looked	0.8278	493/611
reaching	0.6569	184/308	looked away	0.8468	420/423	was distracted by	0.8215	353/369
watching	0.6436	198/376	was looking	0.8425	1167/1564	driving unit 1	0.8113	420/530
looked	0.6432	3495/8297	looking at	0.8347	687/899	when she looked	0.8025	334/381
texting	0.6426	126/104	his phone	0.8341	469/550	looked away from	0.7885	243/176
eyes	0.6347	402/933	her phone	0.8240	356/368	looked down to	0.7881	244/181
bottle	0.6242	134/198	phone and	0.8187	437/540	she was looking	0.7843	313/381
floor	0.6230	170/328	fell asleep	0.8032	689/1024	he looked up	0.7803	263/265
notice	0.6162	336/805	looked back	0.8025	506/712	was looking down	0.7736	230/178
inattentively	0.6158	104/42	eyes off	0.7850	250/212	down at his	0.7677	215/135
tired	0.6034	160/332	at his	0.7791	496/763	he fell asleep	0.7641	304/409
cigarette	0.5946	97/93	attention to	0.7747	273/307	paying attention and	0.7578	246/275
adjust	0.5812	92/111	at her	0.7641	402/620	looking down at	0.7445	195/126
coffee	0.5808	102/164	he fell	0.7532	310/448	she looked down	0.7441	205/174

5.2.4 Selection for DOI, DD, and ID Classifiers

In this section, *distracted and inattentive* versus *neither* classification using the DOI classifier, *distracted* versus *non distracted* classification using the DD classifier, and *inattentive* versus *non inattentive* classification using the ID classifier are performed. A threshold value for probability scores was used to exclude unigrams, bigrams, and trigrams with low probability scores. For example, if a threshold value of 0.50 is set in the U+B approach, then all the unigrams and bigrams having probability scores (by Equation 1 or 3) less than 0.50 are not considered. The best threshold value for each classifier was determined by searching in the range of 0.00 to 0.90 with 0.05 increment. We examined three metrics-Accuracy, AUC and ROC - to find the best threshold. The accuracy value is the ratio of the number of narratives that are correctly classified divided by the total number of narratives. We used top 100 results of each classifier as correctly classified positive cases to find the best thresholds. The ROC is a graph that shows the performance of a classification model over the entire range of sensitivity and specificity, and AUC measures the area underneath the ROC curve. Among the three, we found Accuracy to be the best metric. Therefore, we used Accuracy value as the evaluation metric for comparing the performances of the classifiers.

For the DOI classification, when applied on the 300 manually reviewed dataset described earlier, among U, U+B, and U+B+T, the U+B approach (by both simple Count and weighted count probability Equations) performs the best, and U+B+T performs the worst. We achieved the highest accuracy from the dataset that was prepared by the distracted AND inattentive cases. The U+B+T contains trigrams, and three consecutive words are generally rare, and those found in the training data are unlikely to repeat in the test narratives. For example, “inattentive in his” is a trained trigram with 0.93 probability, but the possibility of these exact three words appearing in a test narrative in the same order is low. The test narrative may have trigrams like “inattentive in her”, “inattentive on his”, “inattentive because his”. For this reason, the U+B+T approach does not work well as expected, but with a large enough training dataset, it may perform better. With the Equation of simple count probability, we achieved the best result using the U+B approach with a 0.35 cutoff value. With the Equation of weighted count probability, we achieved the best result using the U+B approach with a 0.5 cutoff value (Figure 5-9).

Next, the U+B approach using both Equations were applied to the 2020 and 2021 NDOI test set. As mentioned earlier, we used top 100 results of 2020 that were manually reviewed to check the performance of classifiers in the training set. To investigate the performances in the unseen dataset, the top 300 results of 2021 of all the classifiers were manually reviewed and verified. The breakdown of the top 300 manual reviews of 2021 using weighted count probability is shown in Figure 5-9. The accuracy of classifying DOI cases from NDOI cases is 78% using the weighted count probability. Figure 5-9 shows that the DOI cases have a consistent rate in the top 300 results. Considering novelty and brevity, only graphs for Equation 3 are presented in the following sections.

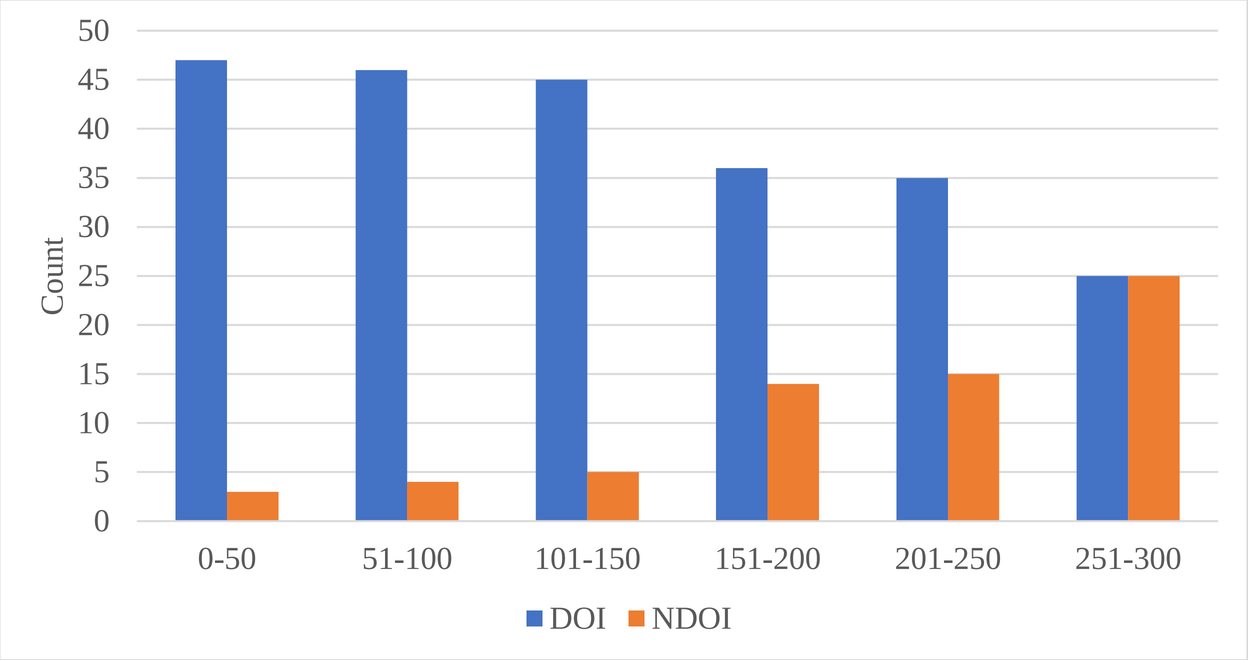

Figure 5-9 DOI Classifier Using the U+B Approach (Eq. 3, 0.50 Cutoff Threshold)

From the Accuracy values obtained on the 500 manually reviewed dataset (described earlier) for the DD classifier, among U, U+B, and U+B+T, U+B approach for both simple count probability and weighted count probability perform the best and U+B+T performs the worst. The reason behind that, as was for the DOI classifier, is that U+B+T contains trigrams, and three consecutive words found in training data are unlikely to repeat in the test narratives. With the Equation of simple count probability, we achieved the best result using the U+B approach having a 0.75 cutoff value. With the Equation of weighted count probability, we achieved the best result using the U+B approach having a 0.65 cutoff value.

Next, this U+B approach using both Equations are applied to the 2020 and 2021 NDD dataset and the top 100 results of 2020 and 300 result of 2021 were manually reviewed and verified. The result's breakdown of the top 300 manual reviews using weighted count probability is shown in Figure 5-10. From Figure 5-10, the Equation of weighted count probability is good at finding distracted cases in the top 300 results of the DD classifier. The accuracy of classifying DD cases from NDD cases is 53.33%. It also shows a consistent rate of DD cases in the top results.

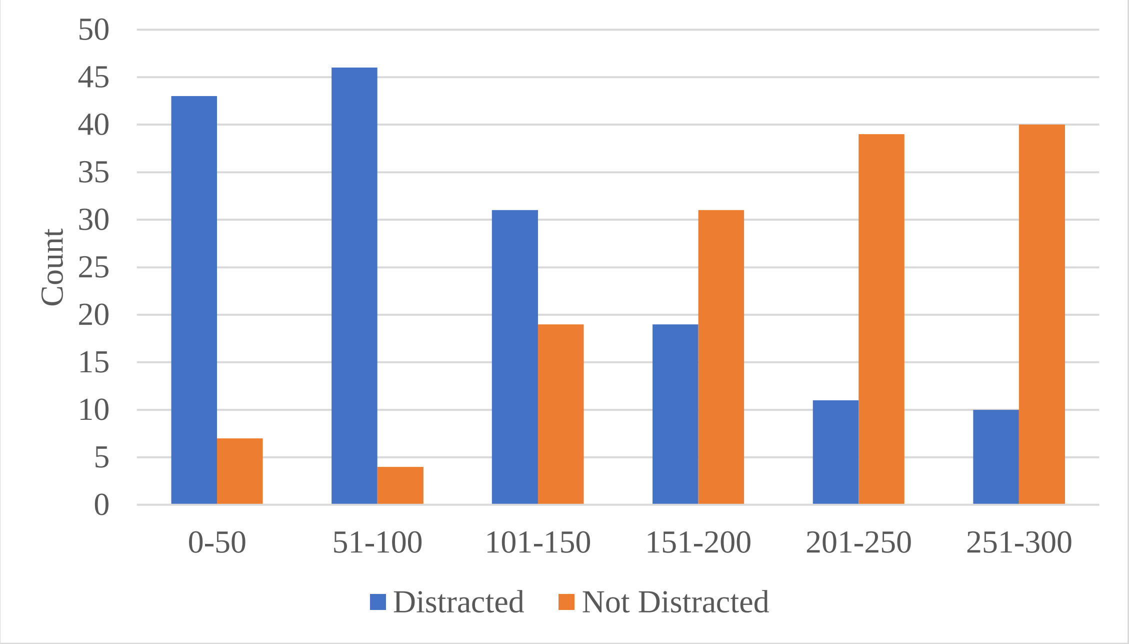

Figure 5-10 DD Classifier Using the U+B Approach (Eq. 3, 0.65 Cutoff Threshold)

From the Accuracy values obtained on the 500 manually reviewed dataset of the ID classifier, among Unigram, U+B, and U+B+T, U (by simple count probability) and U+B (by weighted count probability) approach perform the best, and U+B+T perform the worst. The reason is the same as was for the DOI and DD classifiers, trigrams are rare. With the Equation of simple count probability, we achieved the best result using the U approach having a 0.45 cutoff value. With the Equation of weighted count probability, we achieved the best result using the U+B approach having a 0.75 cutoff value.

Next, these U and U+B approaches using both Equations are applied to the 2020 and 2021 NID dataset and the top 100 results of 2020 and top 300 result of 2021 were manually reviewed and verified. The result's breakdown of the top 300 manual reviews using weighted count probability is shown in Figure 5-11. From this figure, we can see that the Equation of weighted count probability is good at finding inattentive cases in the top 300 results of the ID classifier. The accuracy of classifying ID cases from NID cases in 2021 dataset is 63.33%. However, there is no consistent rate of ID cases in the top results.

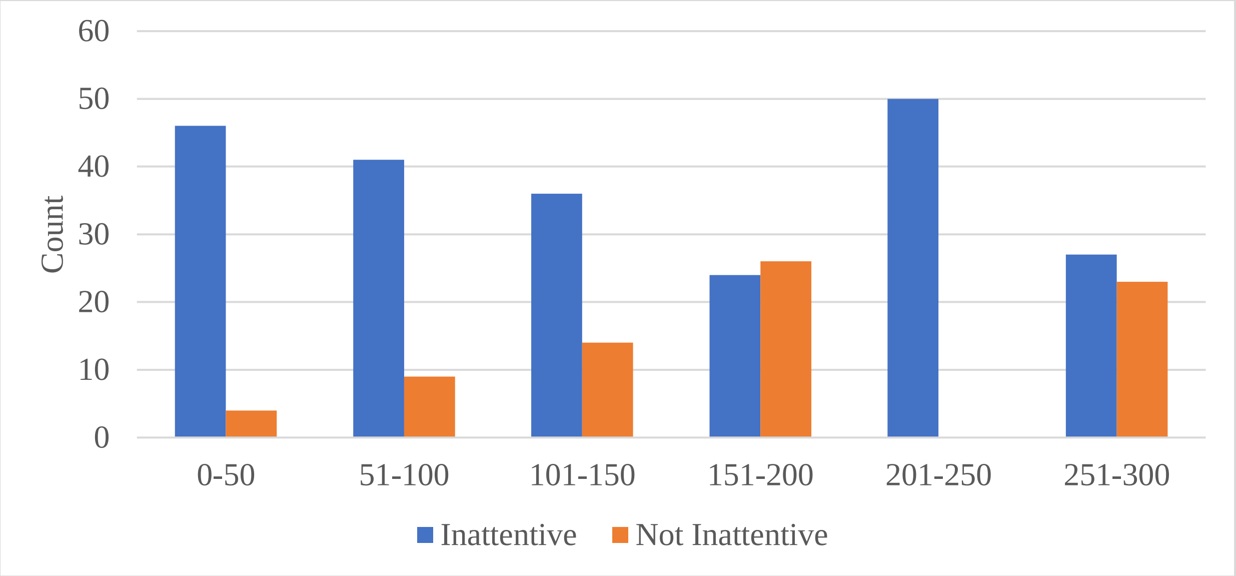

Figure 5-11 ID Classifier Using the U+B Approach (Eq. 3, 0.75 Cutoff Threshold)

The summary of all three classifier's best results using both the Equation of simple count probability and the Equation of weighted count probability is shown in **Table *5-8* Summary of All Results of Best Approaches**Table 5-8. As discussed earlier, the U+B approach yields better results most of the time. Only in one case, U approach shows more promising results. We can also see that the DOI classifier yields very high accuracy values, followed by the ID classifier and finally, by the DD classifier. In general, the DOI classifier performs the best to separate DOI cases from NDOI cases, where the DD classifier performs the worst to separate DD cases from NDD cases.

On the other hand, the ID classifier shows satisfactory performance. In the top results, the ID classifier offers the best performance, followed by the DOI classifier and finally, by the DD classifier. In general, the DD classifier performs the worst in every sector, which also degrades the overall performance on the DD versus ID classification. As a cascaded approach is used to classify cases, every classifier needs to perform well to achieve an overall good result.

Table 5-8 Summary of All Results of Best Approaches

Classifier	Best Approach	Equation	Threshold	Accuracy*	Top 100 results of entire-2020 dataset	% of top 300 results of entire-2021 dataset
DOI	U+B	Simple count probability	0.35	78	89	91.00
	U+B	weighted count probability	0.5	78	91	78.00
DD	U+B	Simple count probability	0.75	65	84	51.00
	U+B	weighted count probability	0.65	54	57	53.33
ID	Unigram	Simple count probability	0.45	45	70	82.33
	U+B	weighted count probability	0.75	49	98	63.33

*: Top 100 results of random sample (300 for DOI, 500 for DD and ID) from 2018-2019 dataset to find the best threshold.

5.2.5 Models to Classify DD, ID and Neither

In the previous sections, we have performed *distracted and inattentive* versus *neither* classification using the DOI classifier, *distracted* versus *non-distracted* classification using the DD classifier, and *inattentive* versus *non-inattentive* classification using the ID classifier. The purpose of this section is to perform all the classifications using the three cascaded classifiers in a single model that are shown in Figure 4-1 to Figure 4-3. The test set of this evaluation includes 270 manually reviewed random cases from 2018-2020 dataset where 93 are distracted (DD), 87 are inattentive (ID), and 90 are Neither (NDOI) Narratives. These models can distinguish inattentive, distracted and neither narrative, which can help the transportation authorities to improve the road safety protocols and the car manufacturers to develop and modify car safety features. We used accuracy which is a ratio of correctly predicted data (narratives) to the total data as the performance metric to evaluate all the models. Figure 5-12 and Figure 5-13 present the accuracies of the DOI, DD and ID classifiers in the models. The overall accuracy of the models can be calculated by taking the weighted average of accuracies of the DOI, DD and ID classifiers.

The best approaches of all three classifiers using both Equations are selected in each of our frameworks/models. With the Equation of simple count probability and all the best approaches for respective classifiers, we achieved the highest accuracy of 70.37% (weighted average of the DOI, DD and ID classifiers) from the Hierarchical Model (**Figure *5-12***). With the best threshold value of 0.75, the cascade-classifier misclassified 27 narratives as neither (NDOI) cases, which are either distracted or inattentive. It misclassified 10 cases as DOI, which are Neither (NDOI) cases. Among the input of 87 ID cases, 41 (47.12 %) are correctly classified, where this cascaded classifier correctly classifies 69 out of 93 input of DD cases (74.19 %). The Priority Model - DD was the second-best model that achieved an accuracy of 66.29%. The Priority Model – ID achieved an accuracy of 58.89%. However, the Priority Model – DD and the Priority Model – ID have a threshold value of 0 for the ID and DD classifiers, respectively. As a result, anything that is classified as NDD in the DD classifier of the Priority Model – DD is considered as ID and anything that is classified as NID in the ID classifier of the Priority Model – ID is considered as DD, which means that the last classifiers become unnecessary in the models. In addition, the error cumulates from the DOI to DD to ID classifiers in the Priority Model – DD, and same thing applies for the Priority Model – ID.

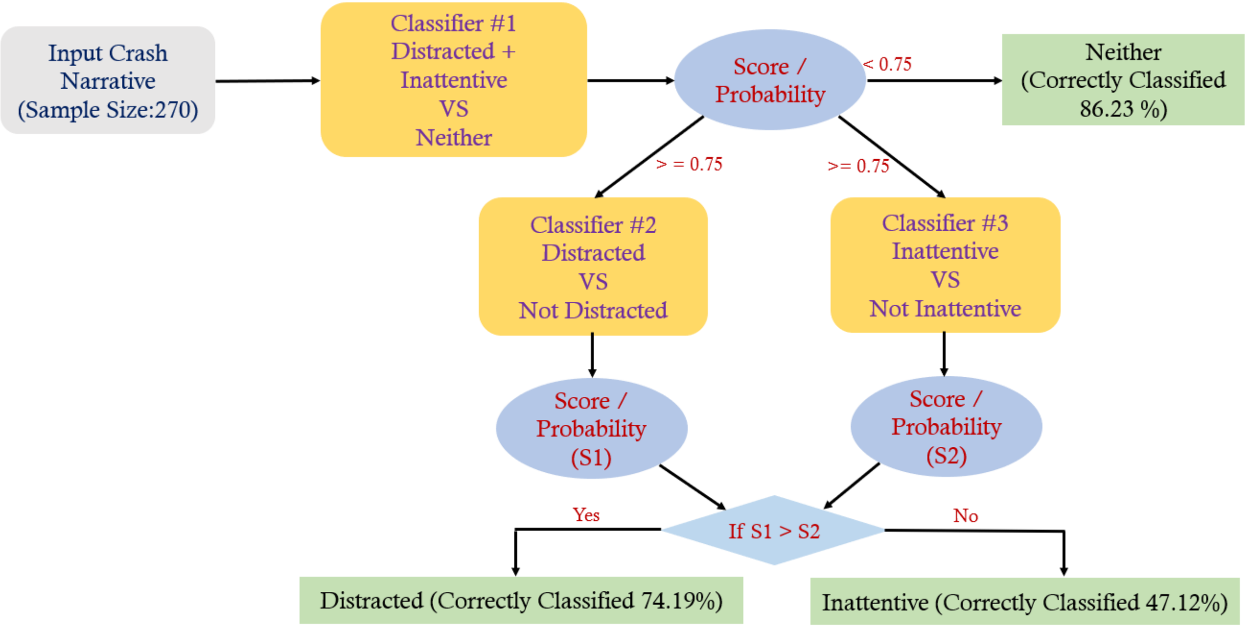
Input Crash Narrative (Sample Size:270)
Classifier #1 Distracted + Inattentive VS Neither
Score / Probability
< 0.75
Neither (Correctly Classified 86.23 %)
> = 0.75
>= 0.75
Classifier #2 Distracted VS Not Distracted
Classifier #3 Inattentive VS Not Inattentive
Score / Probability (S1)
Score / Probability (S2)
Yes
If S1 > S2
No
Distracted (Correctly Classified 74.19%)
Inattentive (Correctly Classified 47.12%)

Figure 5-12 Hierarchy Model for Simple Count Probability Method

With the Equation of weighted count probability and all the best approaches for respective classifiers, with the same test-set including 93 distracted (DD), 87 inattentive (ID) narratives and 90 Neither (NDOI), the highest accuracy of 64.81% is also achieved by the Hierarchical Model (**Figure *5-13***). 83 out of 93 (89.25 %) DD cases are correctly classified with the Hierarchical model, but only 24 out of 87 (27.58 %) ID cases are correctly classified. The cascade-classifier misclassified 15 narratives as neither (NDOI) cases that are either distracted or inattentive and also misclassified 22 cases as DOI that are actually Neither (NDOI) cases. With the Equation of weighted count probability, for the Priority Model-ID (second best model among the weighted count probability), the cascade-classifier predicts 59 narratives as neither (NDOI) cases, which are either distracted or inattentive. With this model, 52 out of 93 DD cases (55.91 %) are correctly classified, where among 87 ID cases, 24 (40.23 %) are correctly classified. The same analysis applies for the Priority Model – DD and the Priority Model – ID for weighted count probability as described for the priority models with simple count probability. For brevity, we did not present graphical analysis for all the models.

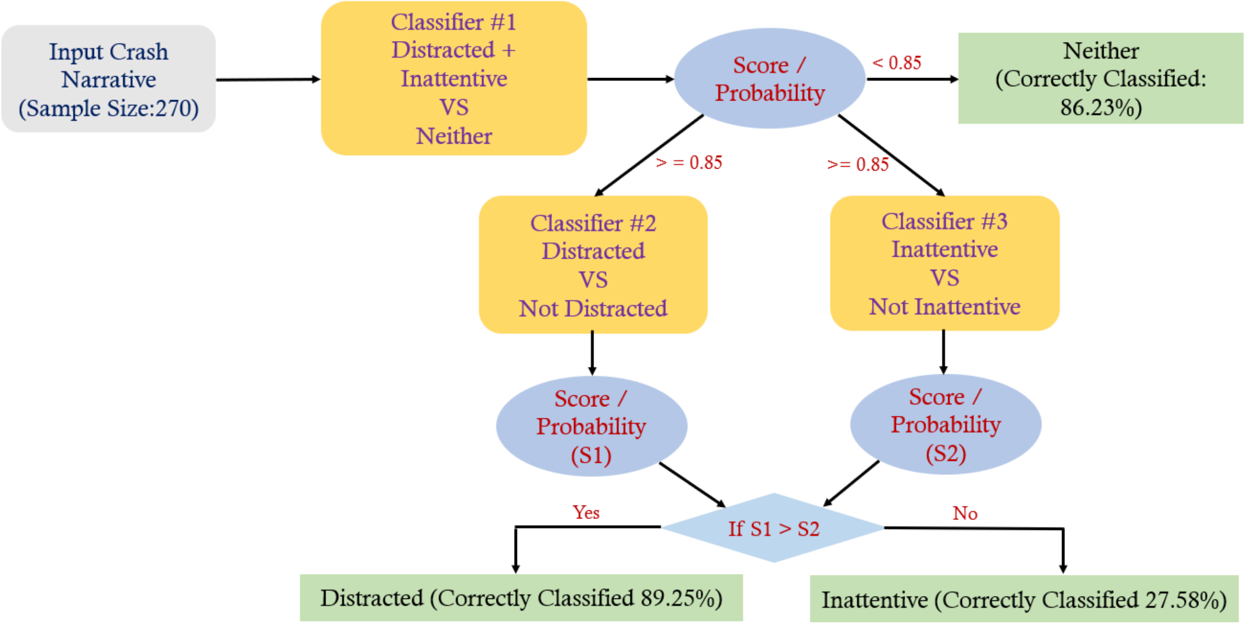
Input Crash Narrative (Sample Size:270)
Classifier #1 Distracted + Inattentive VS Neither
Score / Probability
< 0.85
Neither (Correctly Classified: 86.23%)
> = 0.85
>= 0.85
Classifier #2 Distracted VS Not Distracted
Classifier #3 Inattentive VS Not Inattentive
Score / Probability (S1)
Score / Probability (S2)
If S1 > S2
Yes
No
Distracted (Correctly Classified 89.25%)
Inattentive (Correctly Classified 27.58%)

Figure 5-13 Hierarchy Model for Weighted Count Probability Method

By comparing the result of three classifiers in the models, the ID classifier performs poorly but it can likely be improved with a larger training set (as the training dataset contains very few positive narratives). Also, some police reports did not differentiate distracted and inattentive cases. From this discussion, the Hierarchical model using the simple count probability approach is considered the best model for differentiating distracted and inattentive narratives.

Though we achieved an accuracy of 70.37% for distinguishing distracted and inattentive cases, this research is the first step to automatically separating distracted cases from inattentive cases. The combined DOI and DD classifier in the model performs great on separating distracted and inattentive cases from all other. Also, the modified weighted count probability works very well on finding the most relevant words for each classifier. With an improved dataset for the ID classifier, the results are likely to improve for the cascaded classifiers.

6. CONCLUSION AND RECOMMENDATIONS

This study used probabilistic, NLP and ML techniques to facilitate the identification of certain types of missed crashes from crash narratives. In order to find the best classifiers, multinomial naive bayes (MNB), logistic regression (LGR), support vector machine (SVM), k-nearest neighbor (K-NN), random forest (RF), gated recurrent unit (GRU), and Noisy-OR were tested and compared. As an experimental study, we used the crash narrative of the Wisconsin crash report from 2017 to 2021 to identify missed crashes related to work zone, distracted or inattentive driving, distracted driving only, and inattentive driving only (WZ, DOI, DD and ID crashes).

The data from 2017 to 2018 was used for WZ training, while 2019 data was used for WZ testing. The performance of MNB, K-NN and RF were not satisfactory because the training dataset is too noisy (with about 70% false positives) and has many irrelevant words (words that do not relate to WZ). Although LGR and SVM can detect many WZ-related keywords, their performance in detecting missed WZ crashes was not satisfactory. As the top two performers, GRU and Noisy-OR are comparable in their ability to find missing WZ crashes. The best Noisy-OR result was achieved using both unigram and bigrams from the narratives.

Further analysis on WZ suggests that two types of issues contribute to the misclassification of GRU and Noisy-OR: the mixed phrase that contains at least one highly relevant WZ word (e.g "construction building", “johnson construction ”) and the pesudo-WZ phrase that contains WZ-related words such as “construction barrels”, “construction equipment” and “construction barrier” in a completely irrelevant context. In other words, the narrative with pesudo-WZ phrases contains inadequate information and therefore cannot be classified as a work zone crash.

In addition, Noisy-OR and GRU work differently in noisy narratives and their performance varies by the narrative length. Noisy-OR is more suitable for noisy or lengthy narratives, while GRU is more suitable for less noisy or shorter narratives. In Noisy-OR, an important keyword or positive word can gain a lower probability score (i.e. construction), which leads to a lower classification score for shorter narratives. On the other hand, GRU cannot handle longer narratives that contain many NWZ related words.

Finally, a manual review was conducted for the top 450 cases of Noisy-OR results, and the Noisy-OR recovered 201 missing WZ crashes. The review also indicated that the chance of additional missed WZ crashes beyond the 450 was very low. A follow-up analysis revealed that 73.13% of the missed crashes occurred from 8 a.m. to 6 p.m., with a high percentage happening from 4 p.m. to 5 p.m. A large percentage of those crashes occurred in the summer (July and August) and 43% occurred on urban city streets. The narratives of the cases that have high Noisy-OR scores but are not WZ crashes were carefully reviewed and categorized into the five following groups:

1) Cases with positive words for location or address such as “the Zoo”, “Zoo interchange”: This issue is caused primarily by major roadway construction projects that span multiple years, multiple stages and phases and multiple areas.
2) Cases with positive words for (temporal) traffic control devices such as “concreate barrier”, “median cement”, “attenuator” and “barriers”: Many of these devices, such as median concrete barriers, are permanently deployed to channelize traffic or to protect overpass and underpass structures such as an attenuator near a bridge or at a gore area.
3) Cases with weak positive words for traffic situations such as “congestion” or “backup” which are caused by non-WZ events (i.e., regular congestion or secondary crashes).
4) Cases with strong positive words such as “orange construction cones” or even “construction zone” whose situations are not actually related to a work zone location or work zone activities.
5) Undecided cases, even after a manual review: The authors were conservative and categorized undecided crashes from this study as NWZ crashes.

The location and/or time of a work zone crash can certainly improve WZ classification in types 1 and 5. Such information, however, needs to be linked to and retrieved from a different data source or system such as a lane closure system or a work zone management system. Application of advanced text mining techniques may help improve classification accuracy for cases in types 2 and 3. Unfortunately, no good solutions are available for cases in type 4, but such cases rarely occur. Nevertheless, the discussion underscores the importance of properly documenting the presence of a work zone or work zone activities in the crash narrative.

In the second case study, we applied various forms of words with Noisy-OR to identify DOI, DD, and ID crashes from crash narratives. The 2018 and 2020 datasets were used in training, and the 2021 data was used as a test dataset because of the updates to the distraction data field in 2018. The methods were based on probability scores of unigrams, bigrams, and trigrams and were combined using Noisy-OR. A new and improved way of computing probability scores was introduced to suit the task. The method worked on automatically generated training data that required no manual effort. The classifiers obtained good results despite the noise in the training data. Finally, several methods that combined these classifiers into cascade classifiers to categorize DD versus ID narratives were investigated.

Overall, the DOI classifier with simple count probability method worked well in a new (2021) dataset compared to the DOI classifier with weighted count probability. The threshold value is 0.35 for simple count probability and 0.5 for weighted count probability, respectively. Since a lower threshold value means keeping more keywords for the classifier, the DOI classifier with simple count probability searches more keywords in the narratives compared to that in the classifier of weighted count probability. A DOI narrative that does not have strong DOI related words (high probability scores) can be handled well by the simple count probability method.

The performance of the DD classifier with weighted count probability is consistent in both the known (2018-2020) and new (2021) dataset compared to that of the DD classifier with simple count probability. The threshold value of the DD classifier with weighted count probability is lower than that of the DD classifier with simple count probability. Therefore, a similar conclusion can be drawn for the DD classifier as the DOI classifier with simple count probability – the higher threshold value of 0.65 (compared to that of the DOI classifier) indicates that the DD classifier with weighted count probability will work well in the narrative that has strong DD-related keywords.

The performance of the ID classifier with simple or weighted count probability is not consistent. In the random sample dataset, the accuracy is very low, indicating that the ID classifier does not perform well in the narratives that do not contain strong ID-related keywords. The ID classifier with simple count probability performs well in finding ID cases in the top results of known (e.g., training) and new (e.g., test) dataset. Though the classifier with weighted count probability performed well in the known dataset, it did not perform well in the new dataset; it is possible that the ID dataset is very noisy. During the manual review process, we found that most of the time ID narratives did not contain any inattentive keywords. Moreover, most of the keywords in the inattentive cases were related to distracted driving (Table 5-7), indicating that attentive driving cases were not properly recorded in the structured data.

The performance difference for the DD and ID classifiers in the known and new dataset shows that the new test dataset either has many keywords that are not present in the training dataset or that the keywords are not strongly related to DD and ID cases. The main purpose of this study was to find misclassified DOI, DD and ID crashes from the narratives that are reported as NDOI, NDD and NID, respectively. There is no harm in including a test dataset into the training dataset to generate positive word lists for Noisy-OR.

Compared to the work zone classifier that does not require any threshold value, the DD and ID classifiers perform poorly even in the presence of an optimal threshold. The plausible reason is that the data collector who records the data may not be very familiar with the new data fields in the structured data or may have difficulty distinguishing DD and ID crashes despite the well-defined DD and ID crashes in the manual.

In order to separate DD and ID from the DOI cases, several cascade classifiers (models) have been developed using the DOI, DD and ID classifiers. All cascade classifiers perform well at identifying DOI cases from the narrative. The hierarchical model with both simple count and weighted count probability performed best among all the models. The reason behind this is that after classifying the DOI cases in the first stage of the model, the resulting DOI cases are sent into the DD and ID classifiers. The DD and ID classifiers work parallelly and independently to classify DD and ID cases from the DOI cases, respectively. After the classification task is completed in the DOI classifier, all classified DOI cases are used as inputs for the DD and ID classifiers. In the hierarchical models, the DD and ID classifiers do not classify a narrative as a Neither case; therefore, all the false DOI cases will be classified as either ID or DD. In other words, the performance of the hierarchical model degrades as false DOI cases increase. The number of false DOI cases in the hierarchical model with simple count probability is less than in the hierarchical model with weighted count probability; this is why the hierarchical model with simple count probability worked best in separating DD and ID from DOI cases. The performance of the Priority Model – DD and the Priority Model – ID with simple count probability and weighted count probability is inconsistent. Therefore, we do not recommend using the priority-based models with the presence of at least one poor classifier.

Based on the lessons learned from this study, the following recommendations have been suggested:

1. Shallow machine learning methods such as multinomial naive bayes (MNB), logistic regression (LGR), support vector machine (SVM), k-nearest neighbor (K-NN), and random forest (RF) may not work well for classifying text when the data is noisy and imbalanced.
2. For noisy text data, both Noisy-OR and gated recurrent unit (GRU) can be used, but GRU should be used cautiously.
3. Noisy-OR is the best option when most of the narratives in the dataset are lengthy (e.g, more than 200 words). GRU, however, provides a better result in the opposite circumstance. So, the Noisy-OR is more appropriate for processing imbalanced and noisy crash narratives.
4. GRU is complex, computationally intensive and difficult to interpret. On the contrary, Noisy-OR is very simple, theoritically sound and requires less comoputational power. Therefore, we recommend using both to find the maximum number of missed crashes.
5. Moreover, the accuracy of Noisy-OR is consistant. Such consistency helps to formulate a regression model which can be used to determine the optimam or near optimum number of narratives to review, if required.
6. When a narrative carries information related to two or more crash types (e.g., distracted and inattentive), their individual Noisy-OR classifiers can be used together in a cascading fashion for enhanced results.

Text mining techniques can overcome the limitations of keyword search (i.e., you need to know exactly what you are looking for. Even so, keyword searches often return irrelevant results (false positives) because words often have multiple meanings) and maximize the value of crash narratives by extracting useful and meaningful information. It is anticipated that text mining techniques will play a more and more important role in supplementing structured data fields in crash data analysis.

ACKNOWLEDGEMENT

This study was supported by a grant from the Wisconsin Department of Transportation (WisDOT) Bureau of Transportation Safety (FG-2020-UW-MILWA-05069 and FG-2021-UW-MILWA-05641). Thanks to the University of Wisconsin-Madison Traffic Operations and Safety (TOPS) Laboratory for making crash narrative data available for this research.

REFERENCES

Abay, Kibrom A. 2015. “Investigating the Nature and Impact of Reporting Bias in Road Crash Data.” *Transportation Research Part A: Policy and Practice* 71:31–45. doi: 10.1016/j.tra.2014.11.002.

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.” *ArXiv Preprint ArXiv:1707.02919*.

Amoros, Emmanuelle, Jean Louis Martin, and Bernard Laumon. 2006. “Under-Reporting of Road Crash Casualties in France.” *Accident Analysis and Prevention* 38(4):627–35. doi: 10.1016/j.aap.2005.11.006.

Blackman, Ross, Ashim Kumar Debnath, and Narelle Haworth. 2020. “Understanding Vehicle Crashes in Work Zones: Analysis of Workplace Health and Safety Data as an Alternative to Police-Reported Crash Data in Queensland, Australia.” *Traffic Injury Prevention* 21(3):222–27. doi: 10.1080/15389588.2020.1734190.

Boser, Bernhard E., Vladimir N. Vapnik, and Isabelle M. Guyon. 1992. “Training Algorithm Margin for Optimal Classifiers.” *Perception* 144–52.

Breiman, Leo. 1999. “Randon Forests.” *Machinelearning202.Pbworks.Com* 1–35.

Breiman, LEO. 2001. “Random Forests.” *Random Forests* 1–122. doi: 10.1201/9780367816377-11.

Brindha, S., K. Prabha, and S. Sukumaran. 2016. “A Survey on Classification Techniques for Text Mining.” *ICACCS 2016 - 3rd International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Arround the Globe* 01(i):1–5. doi: 10.1109/ICACCS.2016.7586371.

Chang, Yin-wen, and Chih-jen Lin. 2008. “Feature Ranking Using Linear SVM.” *Feature Ranking Using Linear SVM* 53–64.

Cheng, Yang, Steven Parker, Bin Ran, and David Noyce. 2012. “Enhanced Analysis of Work Zone Safety through Integration of Statewide Crash and Lane Closure System Data.” *Transportation Research Record* (2291):17–25. doi: 10.3141/2291-03.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.” 103–11. doi: 10.3115/v1/w14-4012.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20(3):273–97. doi: 10.1007/BF00994018.

Cuingnet, Rémi, Charlotte Rosso, Marie Chupin, Stéphane Lehéricy, Didier Dormont, Habib Benali, Yves Samson, and Olivier Colliot. 2011. “Spatial Regularization of SVM for the Detection of Diffusion Alterations Associated with Stroke Outcome.” *Medical Image Analysis* 15(5):729–37. doi: 10.1016/j.media.2011.05.007.

Cunningham, Pádraig, and Sarah Jane Delany. 2020. “K-Nearest Neighbour Classifiers 2nd

Edition (with Python Examples).” *ArXiv* (1):1–22.

Daniel, J., K. Dixon, and D. Jared. 2000. “Analysis of Fatal Crashes in Georgia Work Zone.” *Transportation Research Record* (1715):18–23. doi: 10.3141/1715-03.

Das, Subasish, Minh Le, and Boya Dai. 2020. “Application of Machine Learning Tools in Classifying Pedestrian Crash Types: A Case Study.” *Transportation Safety and Environment* 2(2):106–19. doi: 10.1093/tse/tdaa010.

Elias, A. M., and Z. J. Herbsman. 2000. “Risk Analysis Techniques for Safety Evaluation of Highway Work Zones.” *Transportation Research Record* (1715):10–17. doi: 10.3141/1715-02.

Farmer, Charles M. 2003. “Reliability of Police-Reported Information for Determining Crash and Injury Severity.” *Traffic Injury Prevention* 4(1):38–44. doi: 10.1080/15389580309855.

Feldman, Ronen, and Ido Dagan. 1995. “Knowledge Discovery in Textual Databases (KDT).” *International Conference on Knowledge Discovery and Data Mining (KDD)* 112–17. doi: 10.1.1.47.7462.

Fitzpatrick, Cole D., Saritha Rakasi, and Michael A. Knodler. 2017. “An Investigation of the Speeding-Related Crash Designation through Crash Narrative Reviews Sampled via Logistic Regression.” *Accident Analysis and Prevention* 98:57–63. doi: 10.1016/j.aap.2016.09.017.

Gao, Lu, and Hui Wu. 2013. “Verb-Based Text Mining of Road Crash Report.” *Transportation Research Board, 92nd Annual Meeting* 5–16.

Garber, Nicholas J., and Ming Zhao. 2002. “Distribution and Characteristics of Crashes at Different Work Zone Locations in Virginia.” *Transportation Research Record* (1794):19–28. doi: 10.3141/1794-03.

Gers, Felix A., and Fred Cummins. 1999. “Learning t o Forget : Continual Prediction with LSTM.” (470):850–55.

Glen, Stephanie. 2019. “Undersampling and Oversampling in Data Analysis" From StatisticsHowTo.Com: Elementary Statistics for the Rest of Us!” Retrieved July 7, 2020 (https://www.statisticshowto.com/undersampling/).

Graham, Jerry L., and James Migletz. 1983. “Collection of Work-Zone Accident Data.” *Transportation Research Record* 15–18.

Graham, Jerry L., Robert J. Paulsen, and John C. Glennon. 1978. “Accident Analysis of Highway Construction Zones.” *Transportation Research Record* (693):25–32.

Gupta, Vishal, Gurpreet S. Lehal, and others. 2009. “A Survey of Text Mining Techniques and Applications.” *Journal of Emerging Technologies in Web Intelligence* 1(1):60–76.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, T. Labs, and Red Bank. 2013. “Tracking Cellulase Behaviors.” *Biotechnology and Bioengineering* 110(1):fmvi-fmvi. doi: 10.1002/bit.24634.

Hauer, Ezra, and A. S. Hakkert. 1988. “Extent and Some Implications of Incomplete Accident

Reporting." *Transportation Research Record* 1185(January):1–10.

Heidarysafa, Mojtaba, Kamran Kowsari, Laura Barnes, and Donald Brown. 2019. "Analysis of Railway Accidents' Narratives Using Deep Learning." *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018* 1446–53. doi: 10.1109/ICMLA.2018.00235.

Ho, Tin Kam. 1995. "Random Decision Forests." *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* 1:278–82. doi: 10.1109/ICDAR.1995.598994.

Hoel, J., M. Jaffard, and P. Van Elslande. 2010. "Attentional Competition between Tasks and Its Implications." in *European Conference on Human Centred Design for Intelligent Transport Systems, 2nd, 2010, Berlin, Germany*.

Inzalkar, S., and Jai Sharma. 2015. "A Survey on Text Mining-Techniques and Application." *International Journal of Research In Science & Engineering* 24:1–14.

Jeong, Heejin, Youngchan Jang, Patrick J. Bowman, and Neda Masoud. 2018. "Classification of Motor Vehicle Crash Injury Severity: A Hybrid Approach for Imbalanced Data." *Accident Analysis and Prevention* 120(December 2017):250–61. doi: 10.1016/j.aap.2018.08.025.

Kantardzic, Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons.

Khattak, Asad J., Aemal J. Khattak, and Forrest M. Council. 2002. "Effects of Work Zone Presence on Injury and Non-Injury Crashes." *Accident Analysis and Prevention* 34(1):19–29. doi: 10.1016/S0001-4575(00)00099-3.

Korde, Vandana, and C. Namrata Mahender. 2012. "Text Classification and Classifiers: A Survey." *International Journal of Artificial Intelligence & Applications* 3(2):85.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. "Text Classification Algorithms: A Survey." *Information (Switzerland)* 10(4):1–68. doi: 10.3390/info10040150.

Leevy, Joffrey L., Taghi M. Khoshgoftaar, Richard A. Bauder, and Naeem Seliya. 2018. "A Survey on Addressing High-Class Imbalance in Big Data." *Journal of Big Data* 5(1):42. doi: 10.1186/s40537-018-0151-6.

Li, Yingfeng, and Yong Bai. 2009a. "Effectiveness of Temporary Traffic Control Measures in Highway Work Zones." *Safety Science* 47(3):453–58. doi: 10.1016/j.ssci.2008.06.006.

Li, Yingfeng, and Yong Bai. 2009b. "Highway Work Zone Risk Factors and Their Impact on Crash Severity." *Journal of Transportation Engineering* 135(10):694–701. doi: 10.1061/(ASCE)TE.1943-5436.0000055.

Liddy, Elizabeth D. 2001. "Natural Language Processing. In Encyclopedia of Library and Information Science." *Marcel Decker, Inc.* 1–15.

Maheswari, M. Uma, and Dr J. G. R. Sathiaseelan. 2017. "Text Mining: Survey on Techniques and Applications." *Int. J. Sci. Res.* 6(6):45–56.

Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge university press.

Maze, Tom, Garrett Burchett, and Joshua Hochstein. 2005. “Synthesis of Procedures to Forecast and Monitor Work Zone Safety and Mobility Impacts.” *Report*.

McAdams, Rebecca J., Katherine Swidarski, Roxanne M. Clark, Kristin J. Roberts, Jingzhen Yang, and Lara B. Mckenzie. 2018. “Bicycle-Related Injuries among Children Treated in US Emergency Departments, 2006-2015.” *Accident Analysis and Prevention* 118(April):11–17. doi: 10.1016/j.aap.2018.05.019.

Meng, Qiang, Jinxian Weng, and Xiaobo Qu. 2010. “A Probabilistic Quantitative Risk Assessment Model for the Long-Term Work Zone Crashes.” *Accident Analysis and Prevention* 42(6):1866–77. doi: 10.1016/j.aap.2010.05.007.

Nayak, Richi, Noppadol Piyatrapoomi, and Justin Weligamage. 2009. “Application of Text Mining in Analysing Road Crashes for Road Asset Management.” *Engineering Asset Lifecycle Management - Proceedings of the 4th World Congress on Engineering Asset Management, WCEAM 2009* (September):49–58. doi: 10.1007/978-0-85729-320-6_7.

NHTSA. 2010. “Overview of the National Highway Traffic Safety Administration’s Driver Distraction Program.” *NHTSA*. Retrieved May 22, 2021 (https://www.nhtsa.gov/sites/nhtsa.gov/files/811299.pdf).

NHTSA. 2021. “Distracted Driving.” *National Highway Traffic Safety Administration (NHTSA)*. Retrieved May 22, 2021 (https://www.nhtsa.gov/risky-driving/distracted-driving).

Oniśko, Agnieszka, Marek J. Druzdzel, and Hanna Wasyluk. 2001. “Learning Bayesian Network Parameters from Small Data Sets: Application of Noisy-OR Gates.” *International Journal of Approximate Reasoning* 27(2):165–82. doi: 10.1016/S0888-613X(01)00039-1.

R. Dewar and P. Olson. 2007. *Human Factors in Traffic Safety*. 2nd ed. Lawyers & Judges.

Rahman, Md Mahmudur, Lesley Strawderman, Teena Garrison, Deborah Eakin, and Carrick C. Williams. 2017. “Work Zone Sign Design for Increased Driver Compliance and Worker Safety.” *Accident Analysis and Prevention* 106(May):67–75. doi: 10.1016/j.aap.2017.05.023.

Rakotonirainy, Andry, Samantha Chen, Bridie Scott-Parker, Seng Wai Loke, and Shonali Krishnaswamy. 2015. “A Novel Approach to Assessing Road-Curve Crash Severity.” *Journal of Transportation Safety and Security* 7(4):358–75. doi: 10.1080/19439962.2014.959585.

Regan, Michael A., Charlene Hallett, and Craig P. Gordon. 2011. “Driver Distraction and Driver Inattention: Definition, Relationship and Taxonomy.” *Accident Analysis \& Prevention* 43(5):1771–81.

Regan, Michael A., John D. Lee, and Kristie Young. 2008. *Driver Distraction: Theory, Effects, and Mitigation*. CRC press.

Sorock, Gary S., Thomas A. Ranney, and Mark R. Lehto. 1996. “Motor Vehicle Crashes in Roadway Construction Workzones: An Analysis Using Narrative Text from Insurance

Claims." *Accident Analysis and Prevention* 28(1):131–38. doi: 10.1016/0001-4575(95)00055-0.

Trueblood, Amber Brooke, Ashesh Pant, Jisung Kim, Hye Chung Kum, Marcelina Perez, Subasish Das, and Eva Monique Shipp. 2019. "A Semi-Automated Tool for Identifying Agricultural Roadway Crashes in Crash Narratives." *Traffic Injury Prevention* 20(4):413–18. doi: 10.1080/15389588.2019.1599873.

Ullman, Gerald L., Melisa D. Finley, James E. Bryden, Raghavan Srinivasan, and Forrest M. Council. 2008. "Traffic Safety Evaluation of Nighttime and Daytime Work Zones." *Transportation Research Board* (NCHRP Report 627).

Ullman, Gerald L., and Tracy A. Scriba. 2004. "Revisiting the Influence of Crash Report Forms on Work Zone Crash Data." *Transportation Research Record* (1897):180–82. doi: 10.3141/1897-23.

Vomlel, Jiří. 2006. "Noisy-or Classifier." *International Journal of Intelligent Systems* 21(3):381–98. doi: 10.1002/int.20141.

Wang, Jun, Warren E. Hughes, Forrest M. Council, and Jeffrey F. Paniati. 1996. "Investigation of Highway Work Zone Crashes: What We Know and What We Don't Know." *Transportation Research Record* (1529):54–62. doi: 10.3141/1529-07.

Weng, Jinxian, Jia Zheng Zhu, Xuedong Yan, and Zhiyuan Liu. 2016. "Investigation of Work Zone Crash Casualty Patterns Using Association Rules." *Accident Analysis and Prevention* 92:43–52. doi: 10.1016/j.aap.2016.03.017.

Williams, Trefor, and John Betak. 2018. "A Comparison of LSA and LDA for the Analysis of Railroad Accident Text." *Procedia Computer Science* 130:98–102. doi: 10.1016/j.procs.2018.04.017.

Williamson, Ann, A. M. Feyer, N. Stout, T. Driscoll, and H. Usher. 2001. "Use of Narrative Analysis for Comparisons of the Causes of Fatal Accidents in Three Countries: New Zealand, Australia, and the United States." *Injury Prevention* 7(SUPPL. 1). doi: 10.1136/ip.7.suppl_1.i15.

Yang, Yiming. 1999. "An Evaluation of Statistical Approaches to Text Categorization." *Information Retrieval* 1(1–2):69–90. doi: 10.1023/A:1009982220290.

Ye, Fan, and Dominique Lord. 2011. "Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models." *Transportation Research Record* (2241):51–58. doi: 10.3141/2241-06.

Zagorecki, Adam, and Marek Druzdzel. 2004. "An Empirical Study of Probability Elicitation under Noisy-OR Assumption." Pp. 880–86 in *Flairs conference*.

Zhang, Jiansong, Valerian Kwigizile, and Jun-Seok Oh. 2016. "Automated Hazardous Action Classification Using Natural Language Processing and Machine-Learning Techniques2." Pp. 1579–90 in *CICTP 2016*.

Zhang, Xu, Eric Green, Mei Chen, and Reginald R. (Reg) Souleyrette. 2019. "Identifying Secondary Crashes Using Text Mining Techniques." *Journal of Transportation Safety and*

Security 0(0):1–21. doi: 10.1080/19439962.2019.1597795.

Zheng, Dongxi, Madhav V. Chitturi, Andrea R. Bill, and David A. Noyce. 2015. “Analyses of Multiyear Statewide Secondary Crash Data and Automatic Crash Report Reviewing.” *Transportation Research Record* 2514(2514):117–28. doi: 10.3141/2514-13.

Zhong, Botao, Xing Pan, Peter E. D. Love, Jun Sun, and Chanjuan Tao. 2020. “Hazard Analysis: A Deep Learning and Text Mining Framework for Accident Prevention.” *Advanced Engineering Informatics* 46(August):101152. doi: 10.1016/j.aei.2020.101152.