Understanding Crash Severity from Hydroplane Crash Reports using Transformers

Subasish Das[1], Jett Tipsword[1]

Abstract

Safety in transportation is a critical concern, with over a substantial amount of fatalities on U.S. roads. While advancements in safety technologies have improved automobile safety, understanding and preventing accidents require in-depth research into contributing factors. This paper focuses on hydroplaning risks due to extreme rainfall, emphasizing the need for prioritized road design and traffic management. Analyzing crash severity is vital for improving road safety, implementing targeted measures, and developing effective educational programs. The challenge lies in dispersed information within crash reports, making manual examination time-consuming and inconsistent. Natural Language Processing (NLP), particularly Transformer models, offers a consistent and efficient approach. This study employs NLP techniques to classify crash reports, aiming to understand words contributing to severity. Despite inherent limitations, this approach facilitates extensive data analysis, revealing patterns and trends crucial for road safety. Three Transformer models—BERT, RoBERTa, and DistilBERT—are evaluated, considering performance metrics like accuracy, precision, recall, and F1-score. Results show a trade-off between model complexity and accuracy, with DistilBERT demonstrating competitive performance. Insights from Lime explainability tool shed light on significant words influencing predictions. The methodology presents a potential model for similar applications in diverse domains. The study's contribution extends beyond transportation, showcasing Transformer models as versatile tools for understanding and harnessing natural language. Evaluation metrics highlight model strengths and trade-offs, providing valuable insights for model refinement and generalization. This study concludes by emphasizing the importance of tailored interventions based on severity insights. Specific terms associated with severity levels become the basis for targeted measures, such as defensive driving campaigns, treatment awareness programs, and infrastructure improvements. The holistic approach, combining NLP technology and machine learning models, holds promise in significantly reducing accidents and fostering a safer transportation environment.

Keywords:

1. Introduction

Safety in transportation is a major concern for many agencies and policy makers. Every year, over 35,000 people die on U.S. roads from a wide variety of factors (National Highway Traffic Safety Administration , 2016). However, advances in accident safety technologies have significantly improved automobile safety (Uchida et al., 2010). Creating a safer environment for automobile transportation requires appropriate prevention techniques based on a more in depth research of the factor and scenario that lead to the accident.

Extreme rainfall on roads, leads to transient ponds being formed which heighten the hydroplaning risk, with 22% of affected road sections showing an increasing trend in annual hydroplaning occurrences (Salvi and Kumar , 2022). This underscoring the necessity for prioritized hydroplaning road design and traffic management in high-risk regions. Understanding the severity of a hydroplaning based crashes is significant for several reasons. Firstly, it plays a crucial role in improving road safety by identifying patterns and trends that contribute to accidents. Secondly, it helps authorities implement targeted measures, such as traffic regulations or infrastructure improvements, to address specific crash scenarios. Additionally, analyzing crash severity's aids in developing effective educational programs for drivers, promoting awareness and responsible behavior. Finally, analysis and frequency of crash types is vital for insurance companies and policymakers to accurately assess risk factors, determine liability, and formulate policies that enhance overall public safety on the roads.

A significant challenge when working with crash reports lies in the dispersion of essential information regarding crash factors and their involvement within the report itself. For example, details concerning vehicle locations, crash severity, relative speeds, object positioning, and other descriptive elements are often included within the narrative. Furthermore, the manual examination of crash narratives for contributing factors and causes, while valuable, is a time-consuming and costly process due to the varying language used in different reports. Manual review results also lack consistency as they are subject to the unique experiences and judgments of individual reviewers (Qin et

Email addresses: subasish@txstate.edu (Subasish Das), jbt71@txstate.edu (Jett Tipsword)

al., 2021). Therefore, a more consistent and efficient approach for information extraction, such as Natural Language Processing (NLP), becomes imperative.

NLP can quickly yield results for identifying details described in crash reports with reduced bias. Advanced machine learning and deep learning models are proficient in handling abstract documents like crash reports. The pivotal innovation in Transformers lies in the attention mechanism, allowing the model to focus on specific segments of the input sequence when making predictions or generating output (Vaswani et al., 2017). This attention mechanism allows Transformer models to comprehend context, dependencies, and relationships among words or tokens in the text. Transformer-based models excel in parsing the text and identifying specific details crucial for comprehending the incident. These models can break down the narrative into structured information, identifying key elements to determine severity and crash types. Furthermore, they can discern the context and relationships among various pieces of information within the report, which is essential for accurate document classification and a deeper understanding of how crashes unfold post-occurrence.

2. Literature Review

Through the use of analysis and text mining, crashes can be analyzed and used for accident prevention. Many studies have looked into a wide array of methods to analyze crash reports and extract useful information. Fitzpatrick et al. (2017) assessed the accuracy of the "speeding-related" crash designation in traffic safety analyses. Speed is a critical factor in traffic safety, and the researchers aimed to develop logistic regression models to validate this designation by analyzing 604 crash narratives. They study found that only 53.4% of crashes designated as "speeding-related" contained narratives describing speeding as a causative factor. Gao et al. (2013) aimed to extract crucial information from unstructured traffic accident reports. The study employed a verb-based text mining method using NLP techniques to identify main verbs, reconstructing accident sequences. Testing on 945 records from the Missouri State Highway Patrol showed the method's effectiveness in crash classification and understanding accident causes. Hossain et al. (2015) investigated the factors behind truck-involved crashes in Bangladesh. The study compiled a database of 144 fatal truck-related crash reports from online news sources. Various text mining tools were used to identify key contributing factors, including vehicle types, collision types, time of day, driver behavior, and environmental conditions. The outcomes, 'Coming from opposite direction' and 'head-on collision' were highlighted as significant event sequences.

Road accidents, a significant consequence of transportation systems, demand global attention due to their impact on injuries, fatalities, congestion, and economic losses. Berhanu et al. (2023) compared car accidents in low- and high-income countries, exploring predictive techniques such as machine learning and spatial analysis to enhance road safety and alleviate congestion. The research underscores the importance of integrating GIS-based spatial methods and advanced optimization algorithms into road safety planning for accurate prediction models and effective traffic management. Employing spatial analysis techniques and deep learning methodologies enhances the precision of accident prediction models, allowing for informed decision-making and targeted interventions based on factors like road type, driver state, vehicle type, weather, and date to reduce their occurrence. Contributing to safety implementations in tranposrtation, (Chawla et al., 2021) assesses the safety performance of roadside culverts and evaluates the effectiveness of safety treatments to reduce the severity of culvert-involved crashes. Identifying such crashes through police reports, the study links them to the nearest culvert and employs a negative binomial regression model to analyze risk factors. The second stage involves estimating crash costs using the Roadside Safety Analysis Program, assessing scenarios for cost-effective treatments. Findings provide an empirical model for predicting culvert-involved crash risks, suggesting safety grates as a promising option for culverts within the clear zone and recommending guardrails under adverse conditions or when other treatments are not feasible.

Hydroplaning incidents pose a substantial threat to road safety, contributing to a significant number of accidents, injuries, and fatalities. This weather-related phenomenon occurs when water accumulates on road surfaces, resulting in a loss of tire traction and control over the vehicle. Despite its clear impact on road safety, the literature reveals a noticeable gap in comprehensively understanding and classifying hydroplaning incidents. Addressing this gap, Das et al. (2020) illuminate the underreporting of hydroplaning-related crashes in traditional databases, proposing a framework to address this limitation. By employing natural language processing tools on seven years of Louisiana traffic crash data, the study utilizes interpretable machine learning models to classify crash attributes. The eXtreme Gradient Boosting (XGBoost) model emerges as the most effective classifier, emphasizing the importance of interpretability in machine learning models. This research not only reveals trends and precursors in crash data but also suggests the potential for quantitative modeling techniques to address safety concerns. Wang and Ding (2018) analyzed hydroplaning risks during rainy conditions, specifically focusing on truck tires—a less explored area in prior research. Employing three-dimensional fluid-structure interaction models, the study compares hydroplaning speeds across different tire configurations and validates the simulation model using field test results.

The findings reveal superior performance of the wide-base 445 tire over the conventional wide-base 425 tire and dual tire assembly with an 11R22.5 tire in wet conditions. Hydroplaning potential increases with thicker water films on the pavement, and high wheel load or tire inflation pressure positively influences hydroplaning speed. Furthermore, the analysis emphasizes heightened hydroplaning risks for truck tires under sliding conditions compared to free rolling conditions, emphasizing the need for comprehensive consideration when implementing safety improvement measures for driving safety. Expanding the discussion on road safety during adverse weather conditions, Kim et al. (2021) highlights the significance of cautious driving on rainy days, particularly concerning factors such as slippery roads and hydroplaning. The study introduces an artificial neural network (ANN) model to forecast road surface conditions, categorizing them based on friction coefficients as hydroplaning, wet, or moist. Validation of the model's accuracy, employing statistical parameters like Correlation ratio, Mean Squared Error, and Root Mean Square Error, demonstrates high precision in predicting hydroplaning, moist, and wet conditions. The proposed ANN method, when implemented, holds the potential to significantly reduce traffic accidents on rainy days by providing real-time road condition information to drivers through on-board equipment utilizing V2X technology. Building on the understanding of weather-related driving risks, Xiaoduan et al. (2011) contribute insights into the impact of rainfall on driving risk. They emphasize the inadequacy of past research in accurately representing this aspect in highway safety studies. Utilizing radar rainfall data with fine resolution from the National Weather Service, the study assesses crash risk on four highway types over a 4-year period. The findings reveal heightened crash and injury risks during rain, with variations dependent on highway type, location, time of day, crash severity, and characteristics. The fine resolution of the data allows for a more nuanced understanding of risk variations, offering insights for the identification of effective crash countermeasures based on the knowledge acquired.

3. Study Objective and Contribution

The primary goal of this study is to employ NLP techniques, specifically utilizing Transformer model architecture, to classify crash reports. This approach enables a deeper understanding of the words and phrases associated with accidents, particularly those contributing to the severity of the crashes. While it's important to acknowledge the inherent limitations, this approach significantly contributes to research endeavors by facilitating the collection and analysis of extensive crash-related data, which is not restricted by severity analysis (Das , 2021). Researchers can discern crucial patterns and trends, such as geographical hotspots or common driver behaviors implicated in these incidents. This, in turn, provides valuable insights into the influence of various factors on road safety. Armed with this data-driven insight, policymakers can make informed decisions, allocate resources effectively, and design targeted public awareness campaigns, all of which collectively promote safer road behavior and contribute to the overarching goal of transportation safety.

Furthermore, this application showcases the capabilities of Transformer models when it comes to comprehending and processing natural language. Crash reports are often intricate and abstract, with each narrative being distinct due to individual officers' unique perspectives.These narratives are filled with textual information that is clear to human individuals, but is confusing for NLP algorithms and models. Consequently, a specific preprocessing approach is necessary for the Transformer models to excel in parsing and extracting specific information from lengthy and complex narratives. This not only boosts the efficiency of data analysis but also ensures that no critical details go unnoticed. This underscores the prowess of Transformer models in dealing with intricate and context-rich textual data.

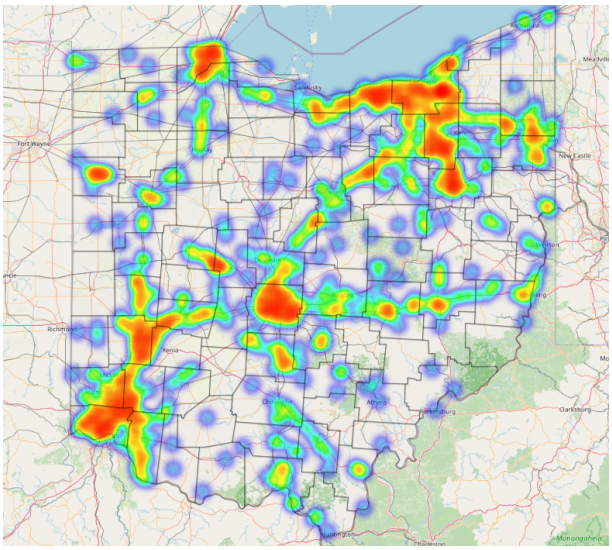
The methodology used for classifying crash reports can potentially serve as a model for similar applications across various domains. Narratives generated by individuals, such as medical records, legal documents, or customer reviews, can all benefit from this Transformer model-driven approach (Jones et al., 2019). By adapting this process to other contexts, valuable insights can be extracted, aiding in making informed decisions based on large volumes of structured and unstructured text. This, in turn, enhances efficiency and effectiveness in multiple industries. Therefore, the utility of Transformer model architecture extends beyond the realm of transportation, offering a versatile tool for comprehending and harnessing natural language in diverse applications.

4. Methodology

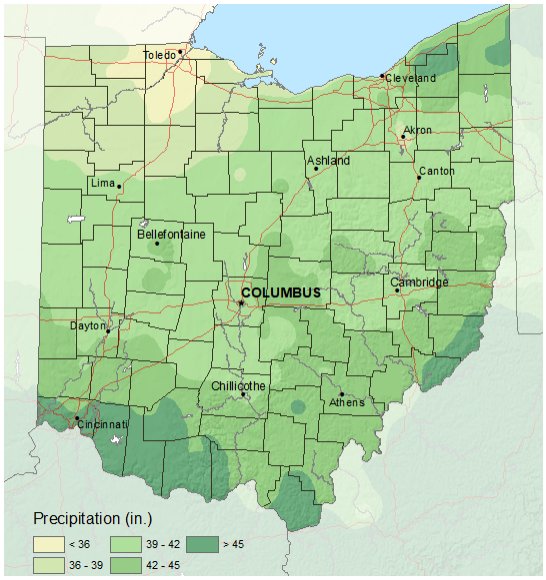
4.1. Data

Two maps of Ohio have been included to show any signs of correlation with how precipitation impacts the number of crashes. Figure 1 (a) is a heatmap where the higher freqeuncy of crashes are shown with warmer colors, were the cooler colors signify lower crash levels. Figure 1 (b), which was provided by PRISM Climate Group at Oregon State University (n.d.), highlights the annual rainfall over the period 1991 - 2020. By comparing this maps

we can see that there is a slight correlation between the southwest and north east sections of Ohio with rainfall and crash frequency. However, this does not imply causation, for Toledo has a high number of crashes with relatively less rainfall than other high crash areas. It may be the case the areas closer to cities experience more crashes than more rural areas due to the increased traffic, however this does not rule out the possibility of higher levels of rain leading to more instances of vehicle hydroplaning.



(a): Target Crash Types



(b): Mean Annual Precipitation (1991 - 2020) PRISM Climate Group, Oregon State University, https://prism.oregonstate.edu, data created 4 Feb 2014, accessed 5 Dec 2024.

Figure 1: Crash Report Heatmap vs Annual Precipitation

Five years (2017-2021) of traffic crash data was obtained from the Ohio Depertment of Transportation. (n.d.). According to the Colorado Department of Transportation. (2019), a crash only needs to be reported if one or more of these conditions are met:

- One or more parties involved were injured or killed because of the crash.
- One or more parties involved did not have insurance.
- One of the drivers specifically requested that a report be completed.
- Any of the vehicles involved incurred $1,000 or more in damages to one or more of the involved parties as a result of the crash.
- Drugs or alcohol suspected of being involved in the crash.

Here the training data is labeled three severity types: O, C, and AB. These three classes all require a crash report to be filled out, as they meet one or more of the conditions for a report to be filed. Consideration for these classes follows Crash Reporting Manual (Colorado Department of Transportation., 2019) guidelines. The three classes are as follows:

- Class 'O' represents situations involving no injuries or exclusively property damage. Within this class, there is a subset specifically indicating incidents without reported injuries, which makes up a minority of the class. This subset is significant as it signifies that these incidents did not result in physical harm or medical consequences for those involved. The primary trigger for categorizing an incident under this class is typically the extent of property damage surpassing a predefined threshold, often set at a specific monetary value, such as over $1,000. This threshold serves as a guideline to ensure that incidents involving significant property damage are formally documented and reported, irrespective of whether injuries occurred. Notably, this class often constitutes the majority of cases in this dataset. This is primarily because accidents leading to property damage are relatively

common. Having a dedicated class for such incidents facilitates the proper documentation and management of property damage claims. Additionally, it aids in evaluating overall road safety, especially considering that even minor accidents can easily exceed financial thresholds like $1,000 in terms of property damage.

  - **Example of Class 'O' Crash Report**: "unit2 was traveling on mockingbird. unit1 was traveling on mockingbird when it went left of center and struck unit2. unit1 stated he was distracted by his radio. unit1 sustained front end damage. unit2 had driver side rear damage. no injuries reported on scene."

- Class 'C' represents non-incapacitating and minor injury. A non-incapacitating injury is one that, while causing discomfort or inconvenience, doesn't prevent an individual from carrying out their usual activities. These injuries can range in severity and typically encompass bruises, minor sprains, superficial cuts, mild burns, low-impact fractures, minor contusions, and mild concussions without loss of consciousness. They typically require minimal medical attention and can often heal with time, rest, or basic first aid. Similarly, minor injuries are generally not life-threatening and don't result in significant impairment. They may involve small cuts, minor burns, sprains, bruises, or simple fractures. While these two types of injuries they may cause temporary discomfort or limit activities briefly, they are not expected to have a lasting impact on a person's overall health or functioning.

  - **Example of Class 'C' Crash Report**: "unit1 traveling ramp. unit1 failed maintain control laid bike causing damage side motorcycle. driver unit1 skidded across ground causing abrasions legs, arms, hands . unit1 refused medical treatment."

- Class 'AB' represent incapacitating/severe and moderate injuries. A severe crash occurs when an individual dies within 30 days of the crash date as a result of injuries sustained in the crash, or is found dead at the scene due to injuries sustained during the crash. Incapacitating injury is defined as bodily injury which, either at the time of injury or a a later time, involves significant risk of death, a risk of substantial permanent disfigurement, a significant risk of loss or impairment of the function of any part of or organ of the body. Incapacitating injury also includes fractures, breaks, or burns to the second or third degree.

  - **Example of Class 'AB' Crash Report**: "due closure smithfield , driver unit1 attempting back smithfield direction . driver unit2 traveling smithfield . driver unit1, struck unit2 front causing damage vehicles. driver unit2 transported hospital treatment . see statements given parties involved. driver unit1 cited ."

Table 1: Yearly Crash Statistics

Year	O	C	AB	Total
2017	84	19	18	121
2018	314	47	60	421
2019	239	43	59	341
2020	364	55	92	511
2021	388	71	97	566
Total	**1389**	**235**	**326**	**1950**

In this research, two datasets were employed—one for training and another for inference. The training dataset is split into two subsets through a 75-25 percent split strategy, dedicated to training and validation purposes. The training set, encompassing 75% of the overall data, plays a pivotal role in instructing machine learning models. Throughout this phase, the models converge on patterns and relationships within the data, enhancing their capability to make accurate predictions or classifications. The validation set, constituting 25% of the data, functions as a tool to evaluate the model's performance during training. This set aids in preventing overfitting by offering insights into the model's generalization capabilities.

The test set is a distinct dataset designated for inference, comprising solely of hydroplaning-related crashes. This dataset was not utilized for training and validation due to its limited size—consisting of only 1950 crash reports—which is deemed inadequate for this application. Instead, a significantly larger crash report dataset that is also from the state of Ohio was utilized for training, employing a transfer learning approach. The utilization of a more extensive dataset in modeling allows the creation of more robust models with enhanced generalization abilities to new data. This approach also facilitates the examination of how the model predicts severity specifically for the targeted category of crashes.

4.2. Data Pre-processing

The pre-processing steps applied to the 'Narrative' column involve a series of text transformations aimed at cleaning and standardizing the data for further analysis. These steps are carried out as follows:

- **Unit Standardization:** First, the script identifies instances of the word unit followed by a number. Unit descriptions can occur in many different ways (unit 1, unit #1, unit1), so the representation has been standardized to ensures consistency in how unit numbers are represented within the text.
- **Location Identifiers Removal:** Common location identifiers like street names, highways, and avenue references are removed to simplify the text and focus on essential details. These do not add much to Transformer based models, making them instances of noise.
- **Stop Word Removal:** Stop word removal for transformer models involves the exclusion of common, non-content words like "and," "the," and "in" from the input text. This process helps reduce noise in the input data, allowing transformer models to focus on more meaningful words and context.
- **Direction Removal:** Next, various patterns related to cardinal directions (e.g., north, south, east, west) and their variations are removed to eliminate unnecessary geographic information that may not be relevant to the analysis. Common Bigrams(west bound, left turn, etc) have also been removed as they are directional specific descriptive words that occur in pairs.
- **Special Characters and Numbers Removal:** Special characters, standalone numbers, and ordinal suffixes (e.g., 1st, 2nd) are removed to further clean the text. Additionally, specific patterns like for highway identification has been removed such as'i-##' where # represents a number.
- **Double Spaces Removal:** The removal of some words and patterns may leave extra spaces behind. Therefore, any consecutive double spaces are replaced with a single space which ensures consistent spacing within the text.

These pre-processing steps are vital for standardizing and cleaning the narrative data within the 'clean_Narrative' column, making it more suitable for subsequent analysis. Table 2 highlights the impact that preprocessing had on the words and characters within the text. Both the word count and character count have been reduced by over 50%, which reaffirms the idea that crash reports are filled with a large amount of noise.

Table 2: Pre-processing Results

Metric	Word Count	Character Count	Avg Word Count	Avg Character Count
Narrative	107,018	575,006	54.88	294.87
Clean Narrative	58,424	363,274	29.96	186.29
Reduction	54.59%	63.18%	54.59%	63.18%

4.3. Transformer Models

The introduction of the Transformer architecture by Vaswani et al. (2017) has significantly transformed the field of Natural Language Processing (NLP) and resulted in remarkable advancements over previous cutting-edge networks. Nevertheless, prior to the emergence of transformer-based models, achieving language processing capabilities comparable to human-level seemed unattainable for computers. This research employs three transformer-based models with identical architecture to investigate their impact on improving classification quality for imbalanced datasets—a common challenge in transportation safety research (Das et al., 2023), (Oliaee et al., 2023), (Weng et al., 2023). These models process input text as tokens, which can be words, word segments, or characters, and convert them into embeddings during the initial layer. Each token is then mapped to the model's internal vocabulary, represented as a sequence of numbers. The models are available in two variations: BASE and LARGE, where the input sequence length limits are 512 and 768 embeddings, respectively. Since the embedding sequence length of our dataset falls below the 512 limits of the BASE version, we exclusively utilized this variant for conducting our tests.

- **BERT**: BERT (Devlin et al., 2019) has emerged as one of the most influential natural language models in contemporary NLP research, playing a pivotal role in popularizing transfer learning in this domain. Its name stands for Bidirectional Encoder Representations from Transformers. Unlike its predecessors, like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which relied on word-level embeddings, BERT utilizes

word pieces, allowing it to handle out-of-vocabulary words more effectively. The model adopts a bidirectional approach to language representation, enabling it to comprehend words, or more precisely, word embeddings, based on the context of the surrounding text. Initially, BERT was trained on two tasks. The first task, mask language modeling (Taylor , 1953), involves masking a portion of text (represented as word embeddings) and training the model to predict the masked word accurately. The second task is the next sentence prediction task, where the model learns to understand the relationship between two sentences. It does this by identifying whether a given sentence is followed by another specific sentence. The successful implementation of these tasks has significantly accelerated NLP research and paved the way for advancements in transfer learning within the field.

- **RoBERTa**: The RoBERTa model (Liu et al., 2019) is an enhanced version of BERT, demonstrating comparable or even superior performance. Its improvements are primarily attributed to the modifications introduced by (Liu et al., 2019) in the BERT training process. These enhancements involve longer training duration, larger batch sizes, and the incorporation of a more extensive range of training data, which includes the CC-NEWS dataset and other English-language corpora. Unlike BERT, RoBERTa is exclusively trained for the masked language modeling objective, employing a dynamically changing masking pattern during the training process. It does not undergo training for next-sentence prediction.

- **DistilBERT**: The DistilBERT represents a significant advancement in the realm of NLP, designed to tackle the challenge of effectively utilizing large pre-trained language models like BERT in situations where computational resources are limited. It introduces a novel approach by incorporating "knowledge distillation" – a process wherein a smaller model, DistilBERT, learns from a larger, more intricate model, such as BERT (Sanh et .al, 2019). To address this challenge, researchers have turned to compressing deep neural networks, resulting in models like DistilBERT. DistilBERT, a distilled (lightweight) variant of BERT, is created through knowledge distillation, a process that trains a smaller model (student) to emulate the behavior of a larger model (teacher) (Silva Barbon and Akabane , 2022). DistilBERT maintains language comprehension capabilities like BERT, but in a smaller, faster format. The distillation process ultimately begets a more streamlined representation of linguistic prowess, substantiated by tangible reductions in model dimensions – approximately 40% – and substantial acceleration in computational processing – approximately 60% (Sanh et .al, 2019). This approach aims to produce a compact model capable of replicating the decisions of the larger model. This is achieved by approximating the distilled model to the function generated by the larger model, which is utilized to classify abundant pseudo data representing attribute values independently. By training the smaller model with pseudo data, the risk of overfitting is reduced, ensuring that the compact model effectively approximates the learned function of the larger model.

4.4. Experimental Settings

During the initial data analysis, a noticeable imbalance in the dataset was found. Therefore, it was necessary to explore augmentation methods and adjust hyper-parameters to improve the model's performance (Oliaee et al., 2023). To fine-tune the models, simulations we conducted with a constant batch size of 32, over eight epochs, with a constant learning rate. For consistent results across all models, the Balanced Categorical Cross Entropy (BCE) (Lin et al., 2018) loss function was used for calculating their loss values. The Adam optimizer (Kingma and Ba, 2017) was used to train the models. The loss values were calculated as follows:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1_{(y_i \in C_j)} w_j \log(p_{\text{model}}[y_i \in C_j]) \tag{1}$$

Where N and C are the number of samples and categories. $p_{\text{model}}[y_i \in C_j]$ represents the model's predicted probability of the sample of index i belonging to the class of index j. The w_j represents the balancing weight that is set inversely to the jth class's frequency.

5. RESULT AND Findings

5.1. Performance Metrics

During the assessment of the models, several performance metrics were taken into consideration. The Accuracy (ACC) represents the portion of correct models' predictions. However, this metric does not discriminate between predictions of the minority or majority class by the models. In datasets with noticeable imbalances, the prediction of the smaller classes affects this metric the least, even though models struggle the most to predict the underrepresented

classes correctly. Precision is considered as a measure of the accuracy of positive predictions and Recall as the measure of the model's ability to identify positive instances correctly. As a result, the macro-averaged F1, the harmonic mean of Precision and Recall, acts as the key performance indicator in the research.

$$\text{Average Precision} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i)} \cdot \frac{1}{C} \tag{2}$$

$$\text{Average Recall} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + TN_i)} \cdot \frac{1}{C} \tag{3}$$

$$\text{Average F1} = \sum_{i=1}^{C} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{4}$$

Where C represents the number of classes. TP_i and TN_i are the correct predictions of samples belonging and not belonging to the ith class. FP_i and FN_i represent the false predictions of samples belonging and not belonging to the ith class. This will be relevant in following sections for analyzing predictive performance of each model.

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

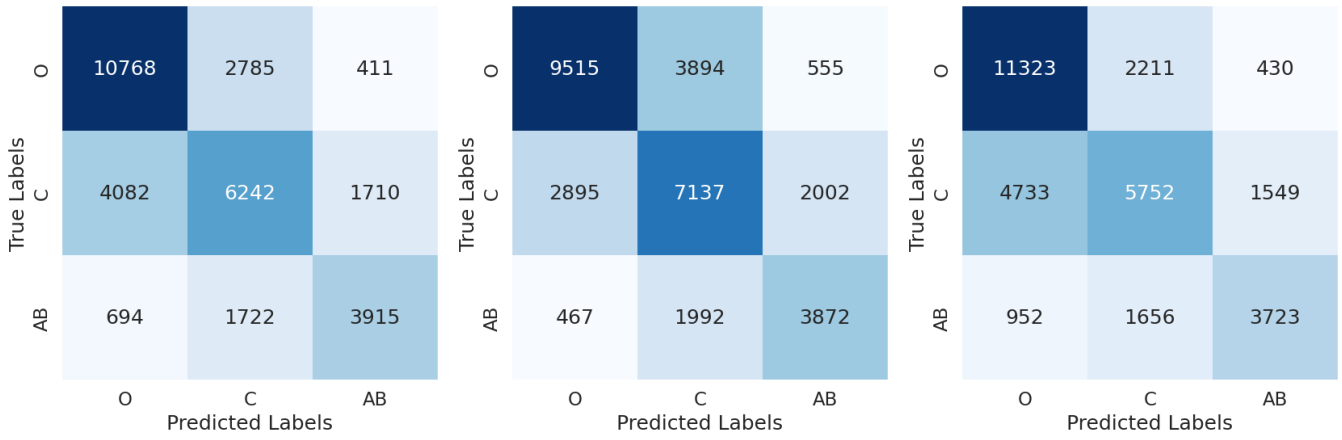
$$FPR = \frac{FP}{FP + TN} \tag{6}$$

$$\tag{7}$$

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

5.2. Results

The tests commenced with a fine-tuning simulation spanning eight epochs, employing a batch size of 32. The fine-tuning process involves iterative adjustments of model weights to enhance classification performance on the training data subset. For this fine-tuning, the process commenced with the use of pretrained model parameters. The loss value was calculated for the training data at each iteration, and the Adam optimizer was employed to update the model parameters. At the conclusion of each iteration, the loss was calculated for the validation data, and a snapshot of the model parameters was saved, along with the corresponding loss and accuracy values. To mitigate over-fitting, where the model becomes specialized in classifying training data at the expense of general performance, we selected the model parameters from the epoch that exhibited the lowest validation loss value.



(a): BERT (b): RoBERTa (c): DistilBERT

Figure 2: Confusion Matrices

The confusion matrices provided in Figure 2 offer a deeper understanding of the observed results across the three different models—BERT, RoBERTa, and DistillBERT—in their classification of crash severity into three distinct

categories. These matrices provide insight of each model's predictive capabilities highlight challenges they face between class assignment. Here we see the confusion matrices for the the models training phase, where the validation set is used for comparing with the truth labels. The values along the x-axis of the matrix represent the labels predicted by the model, while the y-axis represents the target (truth) labels. These matrices highlight the 4 types of predictions TP_i, TN_i, FP_i, and FN_i, where the true labels intersect with the y-axis and the false labels do not intersect with the true class label. A quick way to analyze a confusion matrix is to examine the values along the diagonal, as they indicate the instances correctly predicted by the respective model—where the predicted label aligns with the true label. This technique for model evaluation is useful at identifying the accuracy of the models predictions for each class, especially classes with proportionally fewer samples. However, it is also important to consider the number of values predicted outside of the diagonal, as these highlight the specifc models challenges with predicting precisely.

According to the confusion matrices, BERT stands out as the most accurate model among the three by demonstrating the highest precision in classifying instances across all categories. When looking across the diagonal, this model has 20925 accurate predictions. In particular, its predictions along the diagonal are higher when compared to RoBERTa and DistilBERT. However, even though BERT performs exceptionally well, it is not devoid of errors, with some missclassifications evident in the 'O' and 'C' classes.

RoBERTa, while still offering relatively accurate predictions, exhibits a lower performance overall. When looking across the diagonal where the axis intersect, this model has the fewest correct predictions of 20524. RoBERTa struggled with many of the other classes with fewer samples, but was fairly accurate for the majority classes. The increased focus on majority classes in favor of minority class may raise model accuracy, but this does not indicate optimal performance. These results are significant because RoBERTa is generally regarded as the optimal transformer model for classification tasks (Liu et al., 2019). Most likely, these results highlight the consistent configuration across all three models having a negative impact on RoBERTa. Another explanation could be the dataset itself which is composed of longer sequences than many other NLP tasks due to the narrative of police reports.

DistilBERT, the second-best performer, provides reasonably accurate predictions. When looking across the diagonal, this model has a total of 20524 accurate predictions. Many of the performance across the classes is competitive with the BERT model, with it performing better in class 'O' than the BERT model. Overall, it makes a slightly more number of errors in comparison to BERT, as there is more values outside the diagonal intersection of true and predicted labels. This model may appear to have optimal performance given the high number of correct 'O' class predictions, but this could be a result of the increased predictions for the majority class. Over predicting the majority class may raise accuracy, but lowers precision at discerning between smaller classes. Robust models must be able to discern between minority and majority classes, for a model that only predicts the majority classes correct is not as helpful overall even with increased accuracy.

The relative performance of these models underscores an important trade-off between model complexity and prediction accuracy. Simpler models like DistilBERT, which are computationally less intensive, can often discern between classes for specific tasks such as this one. BERT, with more parameters, also performs well but at a slightly higher level of precision. RoBERTa, while powerful in many natural language processing tasks, is the least accurate and precise of the three on this specific classification task.

Table 3: Performance Metrics After Finetuning

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss
BERT	0.8176	0.2619	0.6473	0.6369
RoBERTa	0.7494	0.3352	0.6348	0.5812
DistilBERT	0.8002	0.2762	0.6433	0.6487

Table 3 provides a comprehensive analysis of the performance metrics for RoBERTa, BERT, and DistilBERT models after fine tuning. Training accuracy is a measure of how well a machine learning model performs on the dataset that it was trained on. It is calculated as the ratio of the number of correctly predicted instances to the total number of instances in the training dataset. A high training accuracy indicates that the model has learned the patterns present in the training data. However, high training accuracy alone does not guarantee good generalization to new, unseen data. Overfitting, where the model memorizes the training data but fails to generalize, is a concern associated with excessively high training accuracy.

Validation accuracy is a metric used to assess the performance of a machine learning model on a set of data that it has not seen during training. It is a measure of how well the model generalizes to new, unseen data. A higher validation accuracy indicates that the model is performing well on new data. It is an essential metric for evaluating the generalization ability of a machine learning model and ensuring that it does not overfit to the training data, which could hinder the performance on the inference data. Here it is clear to see that the BERT had the highest

validation accuracy, with DistilBERT and RoBERTa having slightly worse performance. The ascending order of validation accuracy suggests that DistilBERT outperforms the other models in terms of generalization to unseen data, while RoBERTa exhibits marhinally lower performance in this regard.

The training loss reflects the difference between the predicted output and the true values during training, while the validation loss indicates how well the model generalizes to new, unseen data from training. In the case of BERT, it has a relatively low training loss, suggesting effective learning on the training set. However, the validation loss is higher, indicating that the model might not generalize as well to new data, possibly overfitting. RoBERTa has a higher training loss compared to BERT, but its validation loss is lower. This might suggest that RoBERTa generalizes better to unseen data, even though it's not fitting the training set as closely. DistilBERT falls in between, with a lower training loss than RoBERTa but a higher validation loss than BERT. Balancing these metrics is crucial for selecting a model that not only performs well on the training data but also generalizes effectively to new, unseen data.

Table 4: Inference Classification Reports

	Precision	Recall	F1-Score	
O	0.84	0.85	0.84	
C	0.25	0.51	0.34	
AB	0.65	0.10	0.18	
Macro Avg	0.58	0.49	0.45	
Weighted Avg	0.73	0.69	0.67	
Accuracy				**0.69**

BERT

	Precision	Recall	F1-Score	
O	0.85	0.79	0.82	
C	0.22	0.60	0.33	
AB	0.63	0.07	0.13	
Macro Avg	0.57	0.49	0.42	
Weighted Avg	0.74	0.64	0.64	
Accuracy				**0.64**

RoBERTa

	Precision	Recall	F1-Score	
O	0.82	0.92	0.87	
C	0.29	0.45	0.36	
AB	0.78	0.06	0.12	
Macro Avg	0.63	0.48	0.45	
Weighted Avg	0.75	0.72	0.68	
Accuracy				**0.72**

DistilBERT

Table 4 provides the results from using the fine tuned models to make predictions on the inference hydro planing dataset. Using this type of classification report can highlight how each model preforms with respect to each class. By understanding the models effectiveness on the majority and minory classes, we can assess the models performance. Here, we can see that DIstilBERT had the highest accuracy, with BERT having the second best, and RoBERTa with the worst. This can be attributed to DistilBERT's ability to identify majority class more effectively.

Precision (2) emphasizes the accuracy of positive predictions, with DistilBERT exhibiting a relatively higher precision compared to BERT and RoBERTa. Despite RoBERTa's lower values in other metrics such as validation loss, it showcases a potential strength in handling minority classes. Conversely, both BERT and DistilBERT demonstrate higher precision than RoBERTa, indicating potential challenges in accurately identifying positive instances of the majority classes.

Recall (3), representing the model's ability to correctly identify positive instances, unveils an interesting pattern. RoBERTa exhibits the highest recall among the three models for class 'C', but the lowest for class 'O'. However, this lower recall is compensated by a higher precision, revealing a nuanced trade-off in performance. BERT outperforms RoBERTa in recall, capturing a higher percentage of actual positive instances. DistilBERT, on the other hand, demonstrates the highest recall for class with a great frequency of samples such as class 'O' and 'C', suggesting a relatively better ability to identify class that made up a majority of the training data. This pattern highlights RoBERTa's strength in precision, particularly for minority classes, while BERT and DistilBERT excel in capturing positive instances for the majority class.

The F1-score (4), as a composite metric combining precision and recall, becomes instrumental in assessing a model's performance, especially in scenarios with imbalanced datasets. RoBERTa achieves the lowest F1-score, likely due to the fact that it performs poorly on a majority of the data. This information suggests that the RoBERTa model may benefit from a reduction in the fine tuning process, as it begins to perform worse and possibly underfitting. In

contrast, BERT and DistilBERT outperforms RoBERTa in terms of F1-score, suggesting a relatively better balance between precision and recall. This nuanced analysis underscores the importance of considering minority and majority classes, revealing potential overfitting in BERT and DistilBERT towards the majority class and RoBERTa's strength in handling minority classes.

In classification reports, the macro average provides an equal-weighted mean of metrics for each class, making it suitable for scenarios where all classes are equally important. It is not influenced by class imbalance, so it is useful for when there is significant imbalance such as our dataset. On the other hand, the weighted average considers the contribution of each class based on its dataset proportion, providing a more balanced reflection of overall model performance, especially in the presence of imbalanced datasets. Both macro and weighted averages offer concise summaries, aiding in the assessment of a model's generalization across diverse classes in a classification task.

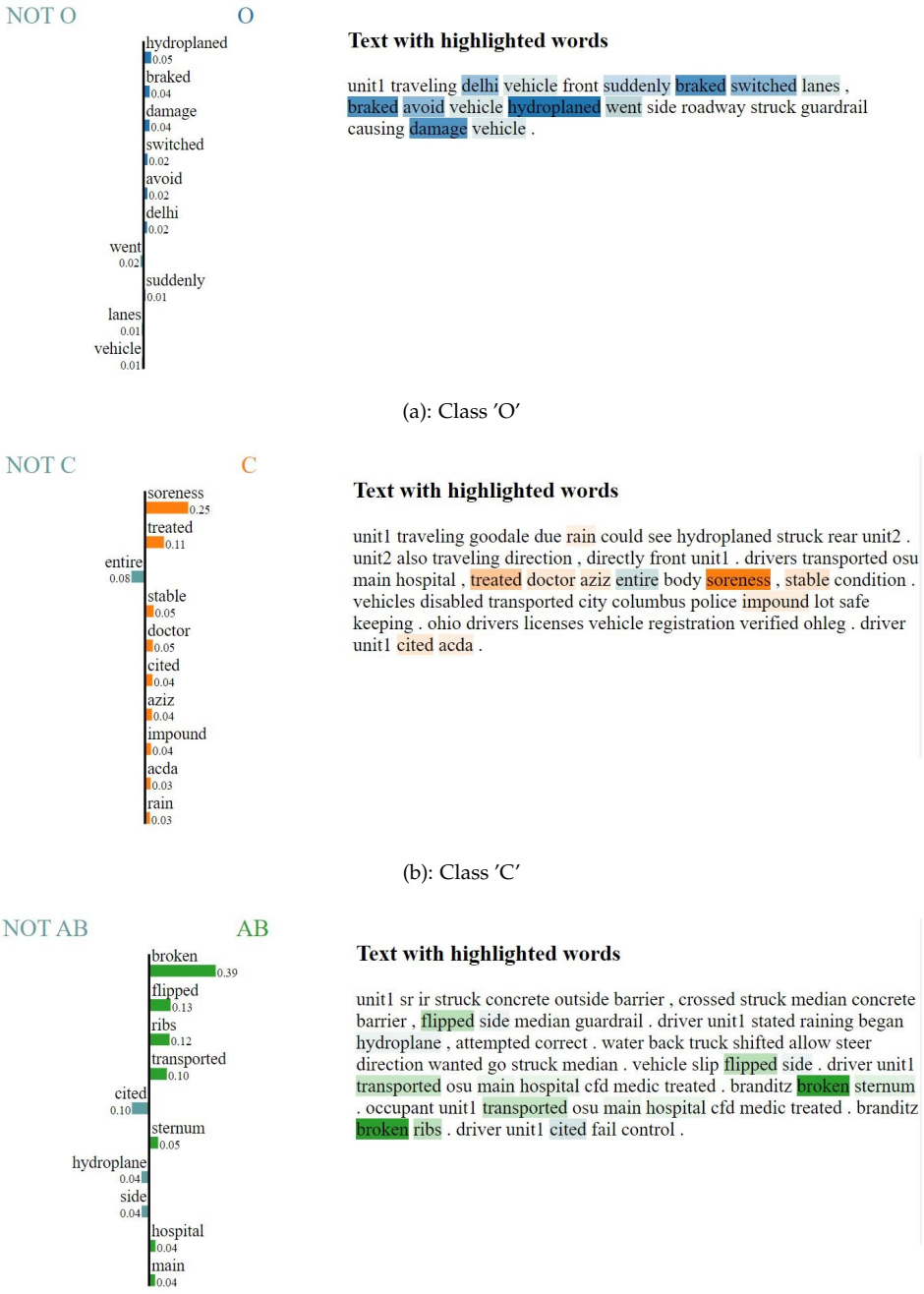
In terms of the macro and weighted averages, DistilBERT again outshines the other models, except in its macro average for recall. These metrics provide a comprehensive summary, considering both the class-specific and overall model performance. Notably, all models face difficulties in correctly classifying the less frequent class AB, suggesting potential areas for improvement. These insights are crucial for model evaluation and can guide further fine-tuning or adjustments in the training process to enhance performance on specific classes and overall model generalization. These insights are crucial for refining models and strategies to prevent hydroplaning accidents in transportation. Additionally, several considerations must be made based on the nature of the dataset and considering the important of each class can help guide a model in focusing on a specific catagory of crash severity.

Figure 3 highlights the words that were considered relevant using the Lime package (Ribeiro , 2023). Lime uses a technique referred to as eXplainable Artificial Intelligence (XAI), a field that is concerned with the development of methods that explain and interpret machine learning models (Linardatos et al., 2020). Machine Learning models experience a trade-off between the performance and their ease of understanding. Deep learning, while a powerful tool, often leads to black-box models that produce high-performance results but are difficult to interpret. In contrast, simpler models like linear (Weisberg, 2005) or decision-tree-based (Safavian and Landgrebe, 1991) models are easier to interpret but less powerful (Farah et al., 2023). This trade-off is important because models that can't be understood are hard to put faith in, especially in fields like healthcare and self-driving cars where ethical concerns arise. To address this issue, there has been a resurgence in the field of XAI (Gunning and Aha , 2019), which focuses on understanding and interpreting AI system behavior. Exactly is the case for this study in which we examine specific words and try to derive their significance from the models classification. Here we have highlighted examples from each class that had the highest probability for that predicted class for the DistillBERT model. Each prediction also has a probability for each class to be predicted, therefore analyzing the examples with the highest probability offers insights into what words the model find significant.

Across the three distinct classes we can observe how each class determines significance to words in the text. The higher the value for each word, the higher the significance to the prediction made by the model. Here we see an example for each class from the DistilBERT models predictions, as it had the highest overall value in the classification report and fine tuning stages. For a majority of the classes, the highest word value correlates to the topic. By understand how machine learning models understand the context of a word for a given prediction we can begin to understand what words contibute the level of severity in a crash report.

When looking at the 'O' class in Firgure 3 (a), we can see the type of words that the model believes to describe a property damage only accident. All reports in the dataset were associated with 'hydroplane' accident types, which the model associated with a less severe crash. Given that the 'O' class had the majority of samples in the dataset, it should come as no suprise that the word 'hydroplane' was identified to be associated with this class. Additionally, several other words have been highlighted such as 'braked', 'switch', and 'avoid'. These words are all is indicative of drivers taking proactive measures to mitigate or prevent collisions. When drivers apply brakes, it suggests a deliberate attempt to slow down or stop, thereby reducing the force and severity of a potential crash. The term "switch" implies a driver's maneuvering, possibly changing lanes or making evasive moves to avoid a collision. Such actions are typically associated with efforts to navigate away from a potential crash, contributing to a less severe outcome. Moreover, the explicit mention of "avoid" signifies a driver's intentional actions to steer clear of a collision altogether. In each case, these words denote a driver's responsive behavior aimed at averting or minimizing the impact of a potential collision, thus aligning with scenarios associated with less severe crashes.

The presence of words like "soreness," "treated," and "stable" in class 'C' descriptions shown in Figure 3 (b) is indicative of the nature and response to moderate or non-incapacitating injuries. The term "soreness" implies discomfort or pain without suggesting severe injuries, often associated with minor conditions that result in temporary discomfort rather than significant impairment. When individuals are mentioned as being "treated," it indicates that they have received medical attention. In the context of moderate injuries, this treatment may involve basic medical care or procedures aimed at addressing less severe health issues at the site of collison. Furthermore, describing an individual as "stable" suggests that their overall medical condition is not rapidly deteriorating. In the context of



(a): Class 'O'

(b): Class 'C'

(c): Class 'AB'

Figure 3: Explaining the predictions using Lime

injuries, "stable" implies that the person is not facing life-threatening conditions and is likely to recover without severe complications. Together, these words collectively convey the level of medical attention and stability associated with injuries of a moderate or non-incapacitating nature.

Figure 3 (c) identifies words such as "broken," "flipped," "ribs," and "transported" in injury-related narratives are strongly indicative of severe, incapacitating, or fatal injuries due to the nature of trauma and damage they imply. When injuries are described as "broken," it signals fractures or breaks in bones, suggesting a significant level of trauma and severity. This term is commonly associated with injuries that involve substantial force or impact, often resulting in a more severe outcome. Additionally, if a vehicle is mentioned as being "flipped," it indicates a forceful event, typically associated with high-speed accidents. The violent nature of such incidents suggests a higher likelihood of severe injuries, contributing to the overall severity of the situation. The inclusion of "ribs" in the narrative is noteworthy, as rib injuries are not only painful but can also lead to complications such as punctured

lungs or damage to internal organs. This adds another layer to the severity of the injuries, indicating a higher level of trauma. Moreover, when individuals are described as being "transported," it signifies the need for urgent medical attention. In the context of severe or fatal injuries, transportation to a medical facility is a crucial step in addressing the gravity of the situation and attempting to provide the necessary medical care. These words collectively convey a narrative of significant trauma, bodily harm, and life-threatening circumstances, pointing towards injuries that are likely severe, incapacitating, or even fatal.

By leveraging the insights gained from the analysis of crash reports through NLP models and XAI techniques, authorities can implement targeted interventions that align with the specific challenges posed by different severity levels. By understanding what descriptive words are the most informative, officers creating crash reports can incrementally implement focus on specific descriptive words. This nuanced approach to road safety not only addresses the complexities of accidents but also maximizes the impact of preventive measures, fostering a safer and more secure transportation environment for all road users.

5.3. Limitations

It's important to acknowledge certain limitations associated with this methodology. Firstly, the effectiveness of the analysis heavily relies on the quality and consistency of the crash reports themselves. Variability in reporting styles, language use, and the level of detail provided by different law enforcement agencies may introduce noise and ambiguity into the data, impacting the accuracy of the NLP models. Moreover, the methodology assumes that the crash reports comprehensively capture all relevant details, which may not always be the case. Essential information regarding contributing factors, environmental conditions, or driver behaviors might be incomplete or omitted, leading to gaps in the analysis. Incomplete or inconsistent reporting can hinder the models' ability to accurately identify patterns and trends, potentially limiting the effectiveness of the proposed preventative measures.

Another limitation lies in the dynamic nature of language and its evolution over time. NLP models, while proficient in parsing and understanding language, may struggle with rapidly changing colloquial terms, new technologies, or emerging trends. This could impact the models' ability to adapt to evolving road safety challenges, requiring continuous updates and refinements to maintain relevance. The reliance on crash reports assumes that all incidents resulting in hydroplaning are accurately classified and reported. Underreporting or misclassification of hydroplaning incidents may lead to an incomplete understanding of the phenomenon, affecting the accuracy of the analysis and subsequent preventive measures.

Additionally, the use of NLP models and XAI techniques introduces an inherent level of complexity and potential bias. The interpretability of these models is still an evolving field, and while XAI methods like LIME provide valuable insights, there may be instances where the models' decisions are challenging to interpret or explain. Ensuring transparency and fairness in the application of these models is crucial to avoid unintended consequences in the formulation of preventative measures.

6. CONCLUSIONS

In conclusion, leveraging Natural Language Processing (NLP) technology for the analysis of crash reports offers a powerful tool for identifying areas of improvement in specific crash type prevention. The utilization of advanced machine learning models, particularly Transformer-based architectures like BERT, RoBERTa, and DistilBERT, facilitates the extraction of meaningful patterns and trends from the narratives. The importance of such analysis becomes evident in tailoring preventative measures for different crash types, with a specific focus on hydroplaning incidents in this study. The integration of NLP models enables a comprehensive understanding of crash severity by dissecting the linguistic nuances within incident narratives. This approach allows authorities to categorize incidents into severity levels and derive actionable insights to enhance road safety. The identified words and phrases associated with different severity levels become the basis for targeted interventions.

In the case of hydroplaning-related crashes, the analysis revealed specific terms such as 'hydroplane,' 'braked,' 'switch,' 'avoid,' 'soreness,' 'treated,' 'stable,' 'broken,' 'flipped,' 'ribs,' and 'transported' as significant indicators. Each term provides valuable contextual information about the nature of the incident, contributing to the overall severity assessment. For instance, terms like 'braked,' 'switch,' and 'avoid' suggest proactive measures by drivers, indicative of less severe crashes. On the other hand, terms like 'broken,' 'flipped,' and 'transported' are strong indicators of severe, incapacitating, or fatal injuries.

The development of educational programs, targeted traffic regulations, and infrastructure improvements can be tailored based on these insights. For less severe crashes, emphasis on defensive driving practices and safety campaigns promoting responsible behaviors can be effective. In cases of moderate injuries, awareness programs about prompt treatment and increased driver awareness may prove beneficial. Severe incidents call for more profound interventions, including enhancements in vehicle safety features, changes in road infrastructure, and

stringent traffic regulations during inclement weather. Furthermore, the analysis of crash reports using NLP models provides a consistent and efficient approach for information extraction, overcoming challenges associated with manual review processes. The findings underscore the trade-offs between model complexity and performance, with simpler models like DistilBERT demonstrating competitive results. The use of interpretability tools like Lime helps shed light on the specific words contributing to model predictions, enhancing transparency and trust in the decision-making process.

In summary, the application of NLP technology, coupled with advanced machine learning models, not only facilitates the analysis of crash reports but also serves as a foundation for implementing targeted and effective preventative measures. This holistic approach to road safety, driven by data-driven insights, holds the potential to significantly reduce the occurrence and severity of accidents, fostering a safer transportation environment for all road users.

References

Colorado Department of Transportation, 2019. Investigating Officer's Crash Reporting Manual. https://www.codot.gov/about/committees/strac/dr3447

Das, S., 2021. Understanding Fatal Crash Reporting Patterns in Bangladeshi Online Media using Text Mining. Transportation Research Record: Journal of the Transportation Research Board 2675, pp-960-971. https://doi.org/10.1177/03611981211014200

Das, S., Dutta, A., Tsapakis, I., 2021. Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study. Transportation Research Record: Journal of the Transportation Research Board 2675, pp-449-462. https://doi.org/10.1177/03611981211002523

Das, S., Datta, S., Zubaidi, H.A., Obaid, I.A., 2021. Applying interpretable machine learning to classify tree and utility pole related crash injury types. IATSS Research 45, 310–316. Available at: https://doi.org/10.1016/j.iatssr.2021.01.001.

Das, S., Vierkant, V., Cruz Gonzalez, J., Kutela, B., Sheykhfard, A., 2023. Bayesian Network for Motorcycle Crash Severity Analysis. Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981231164386

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://doi.org/10.48550/arXiv.1810.04805.

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

Gao, L., Wu, H., Transportation Research Board, 2013. Verb-Based Text Mining of Road Crash Report. p. 12p

Hossain, A., Sun, X., Alam, S., Das, S., Transportation Research Board, 2023. Crash Contributing Factors and Patterns Associated with Fatal Truck-Involved Crashes in Bangladesh: Findings from Text Mining Approach. p. 25p

Jones, S., Fox, C., Gillam, S., Gillam, R.B., 2019. An exploration of automated narrative analysis via machine learning. PLoS One 14, e0224634. Available at: https://doi.org/10.1371/journal.pone.0224634.

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. Available at: https://doi.org/10.48550/arXiv.1412.6980.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal Loss for Dense Object Detection. Available at: https://doi.org/10.48550/arXiv.1708.02002.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel) 23, 18. Available at: https://doi.org/10.3390/e23010018.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Available at: https://doi.org/10.48550/arXiv.1907.11692.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. Available at: https://doi.org/10.48550/arXiv.1310.4546.

National Highway Traffic Safety Administration, USA, 2016. 2015 motor vehicle crashes: overview. Traffic safety facts: research note 2016, 1–9.

Oliaee, A.H., Das, S., Liu, J., Rahman, M.A., 2023. Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. Natural Language Processing Journal 3, 100007. Available at: https://doi.org/10.1016/j.nlp.2023.100007.

Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at the EMNLP 2014, Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. Available at: https://doi.org/10.3115/v1/D14-1162.

Petit, L., Zaki, T., Hsiang, W., Leslie, M.P., Wiznia, D.H., 2020. A review of common motorcycle collision mechanisms of injury. EFORT Open Rev 5, 544–548. https://doi.org/10.1302/2058-5241.5.190090

Berhanu, Y., Alemayehu, E., Schröder, D., 2023. Examining Car Accident Prediction Techniques and Road Traffic Congestion: A Comparative Analysis of Road Safety and Prevention of World Challenges in Low-Income and High-Income Countries. Journal of Advanced Transportation 2023, e6643412. https://doi.org/10.1155/2023/6643412

Das, S., Dutta, A., Dey, K., Jalayer, M., Mudgal, A., 2020. Vehicle involvements in hydroplaning crashes: Applying interpretable machine learning. Transportation Research Interdisciplinary Perspectives 6, 100176. https://doi.org/10.1016/j.trip.2020.100176

Kim, S., Lee, J., Yoon, T., 2021. Road surface conditions forecasting in rainy weather using artificial neural networks. Safety Science 140, 105302. https://doi.org/10.1016/j.ssci.2021.105302

Lopez, D., Malloy, L.C., Arcoleo, K., 2022. Police narrative reports: Do they provide end-users with the data they need to help prevent bicycle crashes? Accident Analysis & Prevention 164. https://doi.org/10.1016/j.aap.2021.106475

Qin, X., Kate, R., Sayed, A., Anisuzzaman, D.M., n.d. Using Text Data from the DT4000 to Enhance Crash Analysis.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Xiaoduan, S., Hu, H., Habib, E., Magri, D., 2011. Quantifying Crash Risk under Inclement Weather with Radar Rainfall Data and Matched-Pair Method. Journal of Transportation Safety & Security 3, pp 1-14.

Uchida, N., Kawakoshi, M., Tagawa, T., Mochida, T., 2010. An investigation of factors contributing to major crash types in Japan based on naturalistic driving data. IATSS Research 34, 22–30. https://doi.org/10.1016/j.iatssr.2010.07.002

Wang, H., Ding, Y., n.d. Prediction of Hydroplaning Risk of Truck on Roadways.

Chawla, H., Megat-Johari, M.-U., Savolainen, P.T., Day, C.M., 2021. Evaluation of Strategies to Mitigate Culvert-Involved Crashes. Transportation Research Record: Journal of the Transportation Research Board 2675, pp 403-417. https://doi.org/10.1177/0361198121992070

Ribeiro, M.T.C., 2023. lime.

National Highway Traffic Safety Administration, USA, 2016. 2015 motor vehicle crashes: overview. Traffic safety facts: research note 2016, 1–9.

Weng, Y., Das, S., Paal, S.G., 2023. Applying Few-Shot Learning in Classifying Pedestrian Crash Typing. Transportation Research Record: Journal of the Transportation Research Board 2677, pp 563-572. https://doi.org/10.1177/03611981231157393

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [WWW Document]. arXiv.org. URL https://arxiv.org/abs/1910.01108v4 (accessed 8.11.23).

Silva Barbon, R., Akabane, A.T., 2022. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. Sensors 22, 8184. https://doi.org/10.3390/s22218184

Berhanu, Y., Alemayehu, E., Schröder, D., 2023. Examining Car Accident Prediction Techniques and Road Traffic Congestion: A Comparative Analysis of Road Safety and Prevention of World Challenges in Low-Income and High-Income Countries. Journal of Advanced Transportation 2023, e6643412. https://doi.org/10.1155/2023/6643412

Weisberg, S., 2005. Applied Linear Regression. John Wiley & Sons.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674. Available at: https://doi.org/10.1109/21.97458.

Gunning, D., Aha, D., 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine 40, 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Farah, L., Murris, J.M., Borget, I., Guilloux, A., Martelli, N.M., Katsahian, S.I.M., 2023. Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know. Mayo Clinic Proceedings: Digital Health 1, 120–138. https://doi.org/10.1016/j.mcpdig.2023.02.004

Taylor, W.L., 1953. "Cloze Procedure": A New Tool for Measuring Readability. Journalism Quarterly 30, 415–433. https://doi.org/10.1177/107769905303000401

Salvi, K.A., Kumar, M., 2022. Rainfall-induced hydroplaning risk over road infrastructure of the continental USA. PLoS One 17, e0272993. https://doi.org/10.1371/journal.pone.0272993

Ohio Department of Transportation, Ohio.gov [WWW Document], n.d.

PRISM Climate Group at Oregon State University [WWW Document], n.d. https://prism.oregonstate.edu/