Topic Modeling using BERTopic with Llama Integration on Crash Narratives of SUV involved Bicycle Crashes

Subasish Das
subasish@txstate.edu

Jett Tipsword
jbt71@txstate.edu

Gian Antariksa
gph27@txstate.edu

Word Count: 7277 words + 3 table(s) × 250 = 8027 words

Submission Date: May 31, 2025

ABSTRACT

With the increasing prevalence of sport utility vehicles (SUVs) on roadways and the growing popularity of cycling as a mode of transportation, understanding the dynamics and factors contributing to SUV-bicycle crashes is important for enhancing road safety. This research employs advanced Language Models (LLMs) to analyze narrative reports of SUV-bicycle crashes, aiming to uncover key insights into the causes, patterns, and potential preventive measures. Preliminary findings suggest that contextual elements, such as road infrastructure, weather conditions, and time of day, play a crucial role in shaping the dynamics of SUV-bicycle collisions. The LLMs aid in identifying the key patterns and correlations within the narratives, shedding light on potential areas for targeted interventions, whether in terms of infrastructure improvement, public awareness campaigns, or policy recommendations. By using the power of advanced LLMs, this study provides a comprehensive analysis that can inform evidence-based strategies aimed at reducing the frequency and severity of SUV-bicycle collisions, ultimately fostering safer and more sustainable urban transportation systems.

Keywords: SUVs, bicyclists, crash, safety, large language models (LLMs), BERTopic, and Llama

INTRODUCTION

Safety in transportation is a major concern for many agencies and policy makers. Nearly half of the 1.19 million annual road traffic crash fatalities worldwide involve vulnerable road users, including cyclists, pedestrians, and motorcyclists (*40*). It is important that everyone on the road, including vulnerable users, are considered when making decisions that impact road safety. Based on an analysis of 49 studies from 13 different countries, Elvik and Mysen (12) concluded that official road crash statistics consistently lacked comprehensive reporting of injuries across all ranges of severity. While injuries sustained by car occupants were commonly recorded, those suffered by cyclists were significantly underreported. In Sweden, 59% of all bicycle collision injuries found their way into official statistics, a trend supported by similar findings in other studies highlighting the systematic underreporting of bicycle-related injuries. The police also exhibited a tendency to overlook collisions involving bicycles and motor vehicles at low speeds, off-road incidents, and single bicycle crashes, as noted by Aultman-Hall and Kaltenecker (3). In France, researchers found that both road user type and third-party involvement strongly influenced underreporting in police crash data. Amoros et al. (2) discovered that in crashes with a third party, cyclists were reported 0.75 times less often than car occupants.

Advances in crash safety technologies can significantly improved safety for all users who share the road (*35*). Creating a safer transportation environment requires effective prevention techniques based on thorough research into the factors and scenarios that lead to crashes. One avenue for improving crash safety technologies is through the analysis of crash reports. These reports, filled out by police officers at crash scenes, provide crucial information for insurance claims, legal proceedings, and statistical analysis of road safety, particularly in response to fatal crashes. However, a major challenge lies in extracting essential information from these reports, such as vehicle locations, crash severity, etc. Manual examination of crash narratives for contributing factors and causes is valuable but time-consuming and costly due to variations in language across reports. Furthermore, the results of manual examination lack consistency as they are subject to the unique experiences and judgments of individual reviewers (*26*). Therefore, it's crucial to utilize more consistent and efficient methods for information extraction, like Natural Language Processing (NLP) and Large Language Models (LLM).

NLP is a branch of AI focused on teaching machines to understand, interpret, and produce human language. It involves developing algorithms and models for processing and analyzing text data, enabling computers to perform tasks like translation, sentiment analysis, and summarization. LLMs, a subset of NLP models, are distinguished by their size and ability to handle large datasets. Built on transformer architectures, these models are pre-trained on massive datasets to capture intricate language patterns and contextual relationships. LLMs advance NLP capabilities by significantly boosting performance across various tasks such as text classification, named entity recognition, and language generation. In essence, LLMs are pivotal in enhancing the effectiveness and sophistication of NLP applications.

The introduction of the Transformer architecture by Vaswani et al. (*37*) revolutionized NLP and LLM, yielding significant advancements over previous networks. This research employs the BERTopic transformer-based model with LLM prompt integration to enhance classification quality for small datasets, a common challenge in transportation safety research (*8*), (*22*), (*39*). However, applying topic modeling to police crash reports faces challenges due to diverse language, requiring careful preprocessing for accuracy and ensuring ethical handling of sensitive information. Defining clear topics is difficult due to the various factors contributing to crashes, demanding ongoing refinement to capture evolving language nuances.

LITERATURE REVIEW

3.1 Under reporting of Bicycle-Related Injuries in Official Road Crash Statistics

Bicycle-related crashes present numerous challenges, particularly regarding the underreporting of injuries in official road crash statistics, leading to gaps in safety evaluations. Shinar et al. (*30*) shed light on the issue by investigating the lack of reports about bicycle crashes in police documentation. Their study, encompassing responses from 7015 adult cyclists in 17 countries, uncovers a mere 10% reporting rate, with significant variation among countries (0.0% to 35.0%). Factors influencing reporting, such as crash type, involved vehicles, and injury severity, underscore the substantial underreporting phenomenon. This highlights the necessity of self-report survey data to comprehensively assess bicycling crash patterns for effective prevention and injury reduction strategies. Building on this, Gildea et al. (*15*) reveal significant underreporting of lower severity cycling collisions and single cyclist collisions in police statistics, introducing biases in available collision data. Their utilization of self-reporting survey data from Ireland emphasizes the importance of nearside-hook, vehicle lane changing, and overtaking maneuvers in cyclist-vehicle collisions, influencing cyclist safety priorities and providing valuable insights for road infrastructural planners, injury biomechanics, and automated vehicle safety. Additionally, Reynolds et al. (*28*) establish the elevated risk of injuries requiring hospitalization for cyclists compared to motor vehicle occupants. To address this, a review of 23 studies focusing on transportation infrastructure's impact on cyclist safety indicates that purpose-built bicycle-specific facilities contribute to reduced crashes and injuries, laying the foundation for initial transportation engineering guidelines for cyclist safety. In summary, these sources collectively highlight the pervasive challenge of underreporting in bicycle-related crashes, emphasizing the need for comprehensive data and effective strategies to enhance cyclist safety and inform road design and policy.

3.2 Role of Advanced Technologies in Enhancing Road Safety Research

Innovations in crash safety technologies, such as NLP and LLMs, hold the promise of significantly improving safety for all road users by enhancing our understanding of crash factors through efficient analysis of crash reports. The utilization of crash narratives has been underscored by Wali et al. (*38*), who, in contrast to traditional quantitative crash data, emphasizes the value of overlooked crash narratives in providing unique contextual information about factors associated with injury outcomes in train-involved collisions. This study seeks to unveil hidden recurring themes and ideas within written descriptions of crashes. Additionally, Lee et al. (*18*) explores the safety performance of automated vehicles (AVs) in mixed traffic, revealing insights from 260 AV collision reports in California (2019-2021). The study identifies factors influencing crash outcomes, emphasizing the need for leveraging crash narrative data to improve AV safety in mixed traffic scenarios. Furthermore, Ghasemi et al. (*14*) assesses urban road safety through a comprehensive approach involving traditional checklists and innovative solutions. By integrating various technologies such as eye trackers, GPS, IMU, OBD2, and video recording, the study analyzes driver behavior and vehicle trajectory, demonstrating that innovative techniques enhance road safety reviews by identifying previously unrecognized hazardous points. This highlights the importance of considering the interaction between drivers and infrastructure in road safety evaluations. Together, these studies underscore the multifaceted approach needed to comprehensively enhance road safety, incorporating both advanced technologies and a nuanced understanding of crash narratives.

3.3 Transformer-Based Models in Natural Language Processing for Transportation Safety

The significant impact of transformer-based models, exemplified by Bidirectional Encoder Representations from Transformers (BERT), extends greatly into the domain of NLP, particularly within the context of transportation safety research. Several studies have highlighted the revolutionary effects of models like BERT on language understanding, emphasizing their role in reshaping the landscape, specifically in the processing of crash narratives and the extraction of valuable insights. Research by Doogan and Buntine (*10*) underscores the significance of evaluating topic models in practical, real-world applications rather than relying solely on conventional metrics. The emphasis on nuanced interpretability aligns with the pursuit of insightful topic prediction in transportation safety research. The incorporation of human evaluations, reflective of applied research, in assessing methodologies parallels the efforts to enhance crash report analysis through the integration of BERTopic with Llama. Building on this foundation, Drosouli et al. (*11*) contribute TMD-BERT, a transformer-based model designed for transportation mode detection utilizing sensor data. This approach leverages attention mechanisms to effectively process entire data sequences, assigning weights to different parts of the input sequence to capture global dependencies. The superior performance of TMD-BERT, achieving a high prediction accuracy of 98.8%, aligns with the broader theme of leveraging transformer-based models for comprehensive improvements in transportation safety. Additionally, the study by Valcamonico et al. (*36*) introduces a framework for road safety analysis using NLP and machine learning, providing an automated classification of accidents for expert analysis. Through a comparative analysis of different textual report representation models and machine learning classifiers, the study identifies the optimal combination of Hierarchical Dirichlet Processes for topic modeling and Random Forests for classification. Applied to a repository of US National Highway Traffic Safety Administration accident reports, the framework achieves a balanced trade-off between classification accuracy and result interpretability. This multifaceted approach to road safety analysis further highlights the potential synergy between advanced language models and machine learning techniques in enhancing our understanding and management of transportation safety challenges.

RESEARCH QUESTIONS

The question of whether researchers can feasibly implement NLP techniques, specifically transformer-based models like BERT, to process and analyze crash reports within the constraints of available resources and expertise, invites a nuanced exploration. The feasibility of employing such advanced NLP techniques hinges on the increasing accessibility of these technologies, which has significantly lowered the barriers to entry for researchers across different levels of technical proficiency and resource availability. Nonetheless, the task is beset with challenges, notably the varied language inherent in crash reports and the imperative of meticulous preprocessing to guarantee analytical precision. Despite these obstacles, the strategic application of NLP to crash narratives holds the promise of profoundly enriching our comprehension of these incidents, thus offering valuable insights to enhance transportation safety research.

This discourse naturally leads to another inquiry: how do recent advancements in NLP and LLMs contribute to extracting meaningful insights from crash narratives, and in what ways can these insights be harnessed to surmount the challenges faced in transportation safety research? The integration of transformer-based models like BERT into the analysis process enables the discernment of complex patterns and contextual relationships within crash reports. Such insights are instrumental in identifying the contributing factors and prevailing challenges in road safety, thereby equipping researchers with the analytical tools necessary to address the existing gaps in crash data reporting and enhance the formulation of preventive measures.

Moreover, the role of BERTopic and LLM prompting in offering fresh perspectives on the classification of thematic clusters within crash reports emerges as a critical consideration. This innovative approach melds topic modeling with the analytical prowess of transformer-based models, facilitating a more profound understanding of crash narratives. By enabling the detection of latent topics and themes, as well as subtle patterns and correlations that might evade traditional analytical methodologies, this advancement heralds new possibilities for research and interventions aimed at bolstering transportation safety. Thus, the integration of NLP techniques, particularly transformer-based models, in analyzing crash reports not only illuminates the path forward for transportation safety research but also exemplifies the symbiotic relationship between technological innovation and its application in solving real-world problems.

STUDY CONTRIBUTION AND OBJECTIVE

The objective of this study is to gain insights into the severity patterns of SUV-bicycle crashes by analyzing a dataset of crash reports. Using advanced NLP techniques, the study seeks to reveal thematic clusters within the dataset, shedding light on the circumstances surrounding these incidents. Specifically, the focus is on understanding the factors contributing to different severity levels, ranging from "Serious injury suspected", "Minor injury suspected" and "Property damage only." Through the application of topic modeling techniques, the research seeks to identify common themes and patterns associated with severe and minor outcomes in SUV-bicycle crashes, encompassing diverse topics that provide possible contributing factors. By revealing the prevalence of specific severity levels within each topic, the study aims to inform targeted safety measures and enhance our understanding of the dynamics influencing the outcomes of these incidents. Furthermore, incorporating LLMs and transformer-based models like BERTopic highlights the evolving use of NLP in transportation safety research, enabling a more streamlined analysis of crash reports and improving the interpretability of generated clusters. This objective aligns with the broader aim of employing state-of-the-art technologies to extract valuable insights from vast and intricate datasets concerning road safety, ultimately aiming to facilitate evidence-based decision-making and the formulation of effective strategies to reduce the severity of SUV-bicycle crashes.

METHODOLOGY

The BERTopic workflow in Figure (1) is structured into five essential stages, each contributing to the comprehensive analysis of textual data. This structured workflow ensures a seamless progression from raw text to meaningful insights, leveraging methods for efficient and accurate analysis of textual content.

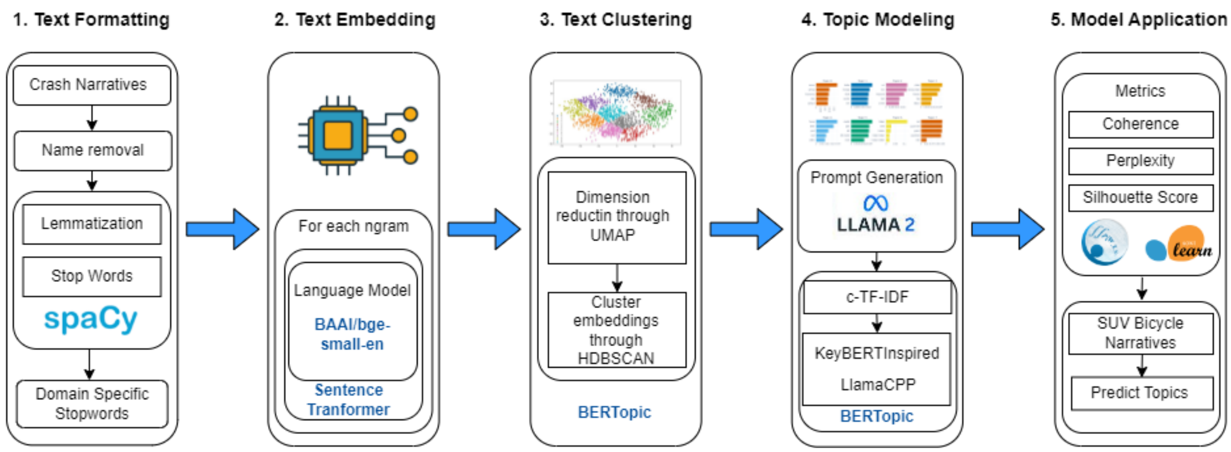

FIGURE 1: BERTopic Workflow

6.1 Benchmarking Topic Models

Figure (2) highlights two performance metrics for BERTopic with llama integration conducted by benchmarking both models using the 20NewsGroup dataset. This dataset, widely used for text classification and clustering tasks, provided a standardized basis for assessing the topic modeling capabilities. The models were evaluated across a range of topic numbers, from 5 to 25, to comprehensively explore their behavior under different configurations.

Figure 2(a) illustrates the relative or "Normalized" values for coherence and perplexity of the BERTopic's predicted topics. These metrics undergo normalization to enable fair comparisons across different numbers of topics. Normalization is imperative in this context due to the inherent variation in scales and ranges of coherence and perplexity scores, which might complicate direct comparisons. By standardizing the metrics, they are brought onto a common scale, typically ranging from 0 to 1, facilitating the assessment of the model's relative performance across varying numbers of topics.

In topic modeling, coherence and perplexity are pivotal evaluation metrics. Coherence measures the interpretability and meaningfulness of the generated topics. A higher coherence score indicates that the words within a topic are more closely related, forming cohesive and contextually meaningful themes. Conversely, perplexity gauges how effectively a probabilistic model predicts a sample of text, typically with lower perplexity values suggesting superior predictive performance and better capture of underlying patterns. However, in the normalization technique used here, perplexity is converted into negative values to maintain consistency in evaluations; higher values still imply better performance. By normalizing these metrics, we ensure that each aspect's contribution to the overall evaluation is appropriately weighted, enabling a more balanced comparison of models with different numbers of topics. This normalization process facilitates the identification of the optimal number of topics that maximizes interpretability while maintaining good predictive performance, thereby enhancing the utility and effectiveness of the topic modeling approach. Based on the normalized metrics, Figure 2(a) shows that there is an inverse relationship between the perplexity and the coherence scores. As the number of topics increases, so does the coherence score. Conversely, as the the number of topics increases, the perplexity score goes down, which in this context shows worsening performance.

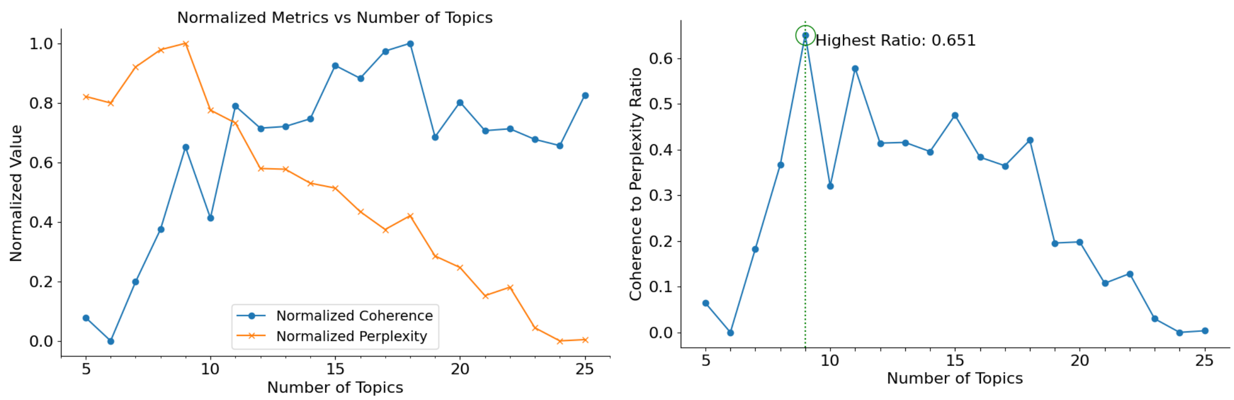

FIGURE 2: BERTopic Benchmarking: Topic Model Performance, (a) Normalized Metrics vs Number of Topics; (b) Coherence to Perplexity Ratio vs Number of Topics

Manually assessing each of these performance metrics independently can yield uncertain results, especially considering their inverse relationship. Hence, converting these normalized metrics into a unified ratio offers a more insightful approach to determining the optimal number of topics. Figure 2(b) demonstrates this methodology by multiplying the coherence and perplexity scores to generate a single score that effectively balances both metrics. According to this figure, the optimal number of topics, based on the performance metric ratio, is 9 topics. This finding is unsurprising, given that the perplexity score for 9 topics is the highest, and the coherence score is also moderately high. Therefore, the subsequent sections will leverage this information by configuring the topic modeling process to generate 9 topics.

6.2 Embedding Narratives

In BERTopic, document embeddings are utilized to create representations in vector space, enabling semantic comparisons among documents with the assumption that those sharing the same topic exhibit semantic similarity. For embedding, BERTopic utilizes the Sentence-BERT (SBERT) framework created by Reimers and Gurevych (*27*), converting sentences and paragraphs into vector representations through pre-trained language models. This approach consistently achieves strong performance across diverse sentence embedding tasks (*33*).The Sentence Transformer model utilized in this process, named "BAAI/bge-small-en" by (*41*), specializes in producing optimal embeddings for sentences and paragraphs. Built upon the SBERT framework, this model has demonstrated outstanding performance across a range of NLP tasks, establishing itself as a reliable option for extracting semantic insights from textual data. These embeddings primarily serve for clustering semantically similar documents, rather than directly contributing to topic generation (*16*). An alternative embedding technique can fit this role if the language models generating documents are fine-tuned to enhance semantic similarity. Consequently, BERTopic's clustering quality improves with advancements in language models, thanks to its adaptability to the latest embedding techniques.

6.3 Clustering Narratives

As noted in other research, when data dimensionality increases, the proximity to the nearest data point tends to approach the distance to the farthest data point (*1*), (*4*). This makes the concept of spatial locality vague, and there's little distinction among different distance measures in high-dimensional space. To address this curse of dimensionality, clustering methods have been proposed to prevent or reduce this conflict Pandove et al. (*23*).Another method involves reducing the dimensionality of embeddings. UMAP, developed by McInnes et al. (*19*), is known for preserving features in high-dimensional data. Therefore, this method of dimensionality reduction is utilized in this study. Importantly, UMAP imposes no restrictions on embedding dimensions, Allowing its usage across language models with differing dimensional spaces. Therefore, we use UMAP to reduce the dimensionality of narrative embeddings generated in Section 4.1.

The reduced embeddings undergo clustering using HDBSCAN (*6*), an extension of DBSCAN (*13*).HDBSCAN detects clusters of varying densities by DBSCAN into a hierarchical clustering algorithm. HDBSCAN employs a soft-clustering strategy, modeling clusters and handling noise by treating it as outliers. This prevents documents that are unrelated from being randomly assigned to clusters, enhancing the representation of topics. Additionally Pealat et al. (*24*) demonstrated that reducing embedding dimensionality with UMAP improves time-series clustering by established algorithms, resulting in enhanced accuracy and faster processing.

6.4 Topic Representation

In BERTopic, we model topic representations based on the documents within each cluster, with each cluster assigned a unique topic. The objective is to understand the distinctiveness of each topic, determined by its cluster-word distribution in comparison to others. To achieve this, the traditional TF-IDF measure, commonly used to assess word importance in a document, is adapted to represent a term's significance to the topic (*16*).

The TF-IDF calculation utilizes two statistics, term frequency, and inverse document frequency (*17*):

$$W_{t,d} = \mathrm{tf}_{t,d} \cdot \log\left(\frac{N}{\mathrm{df}_t}\right)$$

Here, $W_{t,d}$ denotes the weight of term t in document d, representing the Term Frequency (TF), while the inverse document frequency (IDF) measures the information a term contibutes to a document. IDF is calculated by taking the logarithm of the ratio between the total number of documents in a corpus N and the number of documents containing term t.

This process is extended to clusters of documents. Initially, all documents within a cluster are considered as one document by concatenating them. The TF-IDF formula is then modified accordingly for this representation:

$$W_{t,c} = \mathrm{tf}_{t,c} \cdot \log\left(1 + \frac{A}{\mathrm{tf}_t}\right)$$

In this equation, $W_{t,c}$ signifies the weight of term t in class c, where c represents the collection of documents combined into a single document for each cluster. The inverse document frequency is substituted with the inverse class frequency, assessing the information conveyed by a term to a class. This is determined by calculating the logarithm of the average number of words per class A divided by the frequency of term t across all classes, with an addition of one to guarantee only positive values. This class-based TF-IDF approach assesses the significance of words in clusters rather than individual documents, facilitating the creation of topic-word distributions for each cluster of documents Grootendorst (*16*). By progressively combining the class-based TF-IDF representations of the least common topic with its most similar counterpart, the number of topics can be decreased to a user-defined value.

6.5 Large Language Modeling

In LLMs, advanced models have emerged as pivotal tools in understanding and generating human-like text on a massive scale. These models, with their immense capacity for learning from vast datasets, have significantly transformed NLP tasks. This section delves into the landscape of LLM, exploring the methodologies, applications, and impact of these powerful language models on diverse domains.

6.5.1 BERT

Introduced by (*9*), is a significant player in current NLP research, particularly for its role in popularizing transfer learning. By leveraging the contextual understanding provided by BERT, BERTopic is able to create meaningful and coherent clusters, allowing users to identify and explore distinct topics within their text data. Unlike earlier models like Word2Vec (*20*) and GloVe (*25*), which used word-level embeddings, BERT uses word pieces, making it better at handling unfamiliar words. BERT takes a bidirectional approach to language representation, understanding words, or embeddings, based on the context of the surrounding text. Initially trained on two tasks - mask language modeling (*32*) and next sentence prediction - BERT has significantly boosted NLP research, especially in transfer learning.

6.5.2 BERTopic

In early 2022, Grootendorst (*16*) introduced BERTopic, a topic modeling technique leveraging the bidirectional encoder from transformers (BERT). It offers a highly modular and customizable workflow for creating transformer-based topic models. Transformer-based approaches often share a common structure in topic modeling: they initiate by generating document embeddings through a transformer, then proceed to cluster these embeddings to create topics. Following this, a word weighting scheme is applied to extract representative words for each topic. This word weighting scheme called class-based Term Frequency - Inverse Document Frequency (cTF-IDF) was introduced alongside BERTopic. Additionally, BERTopic underwent a comparison with two other transformer-based techniques and two statistical-based techniques, including Latent Dirichlet Allocation (LDA) (*5*), across three different corpora. While BERTopic didn't consistently yield the best topic modeling metrics on every corpus, it maintained competitiveness with other state-of-the-art transformer-based models and consistently outperformed statistical-based models. With its robust metrics and inherent flexibility, BERTopic is well-suited for topic models seeking adaptability as new improvements emerge.

6.5.3 Prompt Models

Prominent prompt based LLMs, such as OpenAI's ChatGPT and GPT-4, represent a notable advancement in artificial intelligence, employing transformer architecture to process input text as tokens and capture contextual relationships. Prompt-based models are language models that generate responses or outputs based on specific input prompts or queries provided by users, allowing them to guide the model's behavior and tailor the generated content to desired themes or topics. The attention mechanism (*37*) in transformers plays a key role in understanding word context within a sentence or document. These models have significantly improved applications like text summarization (*42*). Despite their impact, the detailed methodologies of ChatGPT and its variants remain undisclosed, limiting understanding. Additionally, the high costs associated with API access pose a barrier to widespread adoption across fields, highlighting the evolving landscape and challenges in harnessing the full potential of LLMs Yang et al. (*42*)

In 2023, an new open source model LLM was developed and released by Meta called Llama2 (*34*), which is a collection models, pre-trained and fine-tuned with parameters ranging from 7 billion to 70 billion. Their focus is on optimizing these fine-tuned LLMs, specifically named Llama 2-Chat, for dialogue-related use cases. Through extensive benchmark testing, these models have demonstrated superior performance compared to open-source chat models. Human evaluations, assessing helpfulness and safety, suggest that these models could potentially serve as effective alternatives to closed-source counterparts, such as the GPT family of models. Meta offers a detailed account of their fine-tuning methodology and safety enhancements implemented in Llama 2-Chat, aiming to provide insights and information for the community to contribute responsibly to the ongoing development of LLMs.

Building on the strides made in LLM technologies, exemplified by Meta's Llama2 release in 2023, an integration with the BERTopic library has been introduced. This integration allows BERTopic to leverage the strengths of Llama2 in topic modeling, enhancing adaptability and performance across different corpora (*16*). The collaboration seeks to empower the community in contributing responsibly to the ongoing development of LLMs, addressing barriers such as API costs and promoting accessibility in various fields.The process of topic modeling involves using two representation models: "KeyBERTInspired" and "LlamaCPP". First, KeyBERTInspired extracts keywords that evaluates the significance of words in the narratives, generating indicative keywords. Following this, the LLama LLM utilizes a prompt-based approach to further extract keywords, presenting a topic description alongside documents. The documents here are the preprocessed narratives that describe the crash. Our aim is to elicit concise and meaningful input for labeling topics, with the restriction of a maximum of 5 words per label. The provided prompt includes the keywords and documents for the language model to process:

Q: I have a topic that contains the following documents:

[DOCUMENTS]

The topic is described by the following keywords: '*[KEYWORDS]*'.
Based on the above information, can you give a short label of the topic in at most 5 words?
A:

DATA

Five years (2017-2021) of traffic crash data was obtained from the Ohio Department of Transportation (*21*). According to the Ohio Department of Transportation (*21*), a crash only needs to be reported if one or more of these conditions are met: 1) One or more parties involved were injured or killed because of the crash, 2) One or more parties involved did not have insurance, 3) One of the drivers specifically requested that a report be completed, 4) 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, 5) Drugs or alcohol suspected of being involved in the crash. Here the data is labeled three severity types: "KAB", "C", "O". 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 Ohio Department of Transportation (*21*) guidelines. The three classes are as follows:

- Class "KAB" represent fatal, incapacitating/severe and moderate injuries.
  - **Example of Class "KAB" Crash Report**: "unit #1 was traveling west on tr 0115. unit #2 was traveling south on the ashtabula county metro parks greenway trail and failed to stop at a stop sign. unit #1 struck unit #2. both vehicles traveled off the south side of tr 0115 and the rider of unit #2 was ejected."
- 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.
  - **Example of Class "C" Crash Report**: "unit #1 was stopped on heights ave. waiting to turn right onto northfield rd. north bound. when unit #1 began to make it's right turn unit #2, a bicyclist, was traveling south on the sidewalk along northfield rd. unit #1 was then struck by unit #2 who had the right of way to turn right."
- Class "O" represents situations involving no injuries or exclusively property damage.
  - **Example of Class "O" Crash Report**: "unit 1 was riding a bicycle northbound , on the sidewalk. unit 2 was leaving to turn southbound. units struck on the sidewalk. unable to determine who was at fault."

Table 1 highlights the severity by year for each class of severity on the raw data. This provides a general idea of how the severity's are distributed over time.

Category	**2017**	**2018**	**2019**	**2020**	**2021**	**Total**
Serious Injury Suspected (KAB)	39	35	28	32	45	179
Minor Injury Suspected (C)	256	222	185	143	183	989
Property Damage Only (O)	54	51	48	43	35	231
Total	349	308	261	218	263	1399

TABLE 1: Crash Severity Distribution by Year

7.1 Data Preparation

The pre-processing steps applied to the 'Narrative' column involve a series of text transformations aimed at cleaning and standardizing the data for analysis. These steps are crucial for refining the narrative data within the 'clean_Narrative' column, making it better suited for subsequent analysis. These steps are carried out as follows:

- **Text Cleaning and Standardization:** We streamlined the text by removing common names using lists (*31*), filtering out stop words with the spaCy library, and applying lemmatization to simplify words to their base form. This process includes the elimination of inflectional endings, which helps in standardizing words by their root forms, reducing noise and focusing on meaningful content.
- **Domain-Specific Refinements:** Further, we eliminated domain-specific stop words, directions, location identifiers, and patterns like cardinal directions and common bigrams related to directions to remove irrelevant geographic and contextual details. This step ensures the model concentrates on the most relevant aspects of the narratives.
- **Further Text Simplification:** Special characters, numbers, and any specific patterns indicating locations or directions, such as highway numbers, were also removed. Narratives reduced to less than 10 words were considered to lack sufficient information for analysis and were thus excluded, allowing the model to focus on more informative texts.

RESULTS

Topic Count	Normalized Coherence	Normalized Perplexity	Silhouette Score
8	0.5290	0.9445	0.4206

TABLE 2: BERTopic Evaluation Metrics

Topic modeling, often perceived as subjective and reliant on human judgment, requires human input to validate the usefulness and acceptability of generated topics. However, dismissing topic modeling metrics as useless would be a misconception, as they play a crucial role in providing valuable insights, especially when comparing models with a large number of topics. The combination of human evaluation and topic modeling metrics constitutes an optimal approach to assess the model's performance, as emphasized by Doogan and Buntine (*10*). Table 2 presents the performance metrics discussed earlier, serving as a means to assess the effectiveness of a BERTopic model. An integral aspect of the HDBSCAN clustering algorithm is to detect outlier topics and subsequently group the documents that do not conform with the rest of the topics. As a consequence of this outlier detection process, our model discards the outlier topic, leading to one less topic than initially assigned to the model. Consequently, the total number of topics generated by our model amounts to 9, comprising 8 useful topics and 1 outlier topic.

The coherence of a topic model assesses its ability to identify meaningful relationships among words within topics (*7*). We employ the same normalization technique as described in Section 5.1, with scores ranging from 1 (optimal performance) to 0. The normalized coherence score obtained in our experiment is 0.5290 (coherence score 0.3), indicating a moderate level of interpretability and coherence within the generated topics. While this suggests some success in capturing meaningful word relationships, there is room for improvement to achieve higher coherence levels. Additionally, the perplexity in this experiment is relatively high, with a normalized value close to 1. This suggests a minor level of uncertainty or surprise in the model's predictions.

However, based on previous experiments, it confirms that 9 topics yield the best possible perplexity value for this dataset and modeling configuration.

Clustering quality is assessed using the silhouette score, which yields a value of 0.4206, indicating moderate performance. Silhouettes, as described by (*29*), depict each cluster's compactness and distinctiveness, visually highlighting well-contained objects and those transitioning between clusters. Combining these silhouettes into a single plot offers a comprehensive overview, aiding in cluster quality assessment and data distribution understanding. The average silhouette width serves as a validity measure, helping determine the optimal number of clusters. Scores above 0.5 suggest good clustering, while those below 0.25 indicate poor clustering, and scores between 0.25 and 0.5 denote fair clustering. However, evaluating clustering algorithms goes beyond this metric, considering factors such as cluster number, size, shape, and domain-specific knowledge for a comprehensive effectiveness assessment.

Overall, the BERTopic model demonstrates moderately coherent topics, potentially effective predictive performance (considering perplexity), and moderately well-defined clusters. Interpretation should be contextualized based on the specific requirements and objectives of the topic modeling application. Other studies comparing performance metrics should use raw scores for comparison, as normalized values are relative to benchmarking results. The topic clusters shown in Figure 3 result from HDBSCAN for document clustering, UMAP for dimensionality reduction, and labels from the BERTopic and Llama topic representation model. HDBSCAN is effective in identifying clusters of varying shapes and densities, revealing specific patterns within the document dataset. It also detects potential outliers or noise points that may not conform to traditional cluster structures. After applying HDBSCAN, high-dimensional BERT embeddings are reduced to a two-dimensional space using UMAP, which emphasizes both local and global structure. The cosine metric is selected for distance calculations in UMAP due to its suitability for high-dimensional data like word embeddings.

To enhance the interpretability of the resulting clusters, labels generated by the LLM model are integrated and assigned to the topics created by BERTopic. This integration allows for the structured topics produced by LLM to annotate and categorize the clusters in the visual representation, contributing to a more nuanced understanding of the document content. The resulting reduced embeddings represent documents in a two-dimensional space, where each point corresponds to a document. Incorporating LLM labels enhances the interpretability of the clusters, aiding in the identification of patterns, similarities, and differences among documents. The visual representation, generated using these reduced embeddings and integrated LLM labels, offers an insightful and visually appealing portrayal of the document clusters. Additionally, the HDBSCAN algorithm aids in identifying outliers, which may represent documents that deviate significantly from the main clusters, providing valuable insights into potential anomalies within the dataset. This combined approach provides a richer context for interpretation, enabling a more comprehensive exploration of the content and patterns within the SUV Bicycle Crash Report dataset. The topic clusters provide insight into the relative semantic similarity between topics, with similar topics being group together. This approach proves particularly beneficial for identifying trends in severity related to bicycle crashes. By observing how topics are grouped together, we gain insights into common themes, patterns, and recurring issues within the dataset.

Consider the proximity of "Cyclist Injuries," "Vehicle Crashes with Bicycles," and "Driveway Crashes" within the cluster. This close grouping suggests significant semantic similarities among these topics, particularly regarding instances where a vehicle collides with a cyclist. By

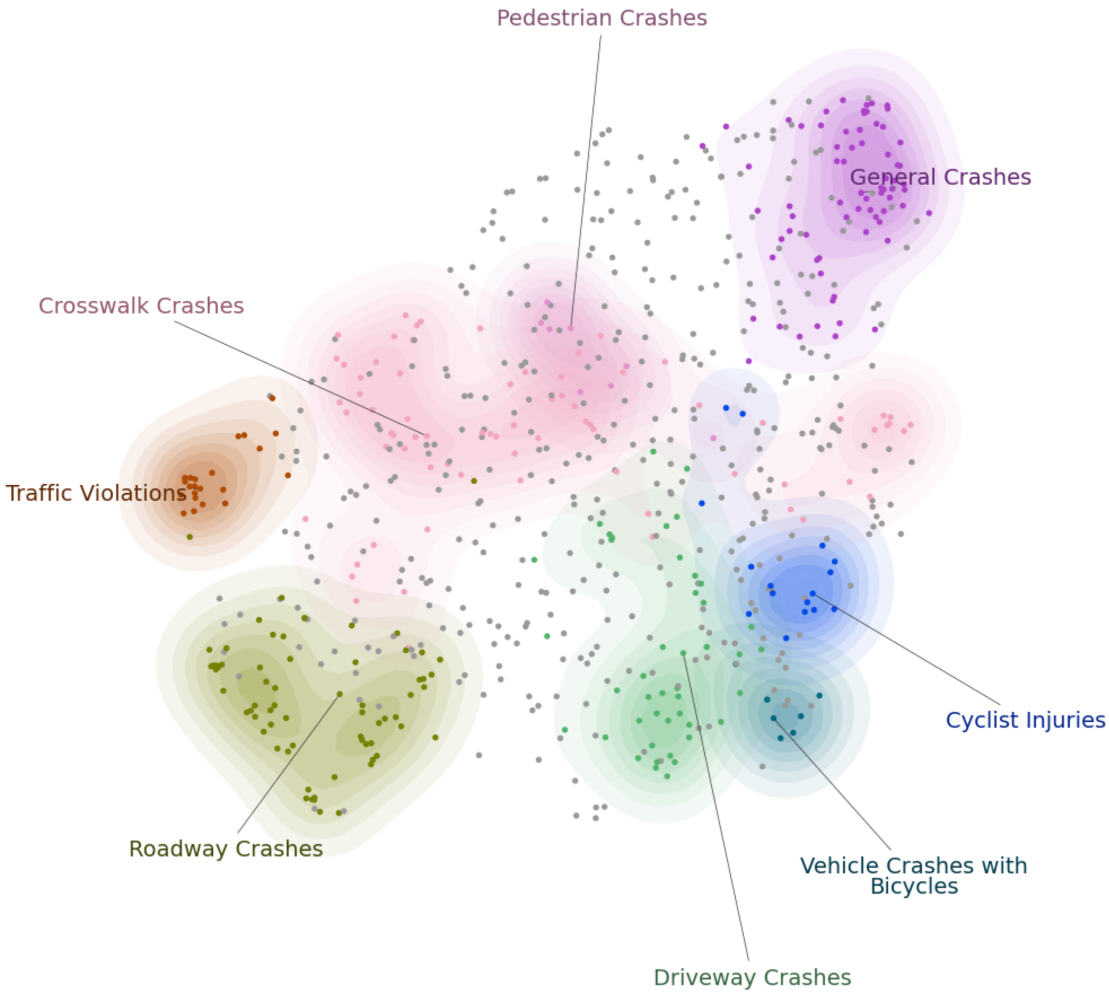

FIGURE 3: Topic Clusters Visualized

clustering these topics together, we can effectively identify and explore recurring themes and issues related to cyclist safety and vehicular interactions. Another instance of semantic similarity among grouped topics is evident with "Crosswalk Crashes" and "Pedestrian Crashes," both representing specific locations where interactions between vehicles and bicycles/pedestrian frequently occur. These topics describe scenarios within a roadway setting, allowing for targeted analysis of particular areas of the road such as crosswalks. Consequently, the content within these topics may share similar keywords, contextual elements, or severity levels, indicating common characteristics across various topics. Furthermore, crash reports with limited information are filtered out as outliers and are not colored. This filtering process is essential to ensure that the results are not skewed by uninformative topic descriptions. By excluding such outliers, we can focus on more informative topic clusters, thereby enhancing the quality and reliability of the insights derived from the analysis.

Table 3 provides a comprehensive overview of the relative severity levels within distinct topics related to SUV-bicycle incidents. Examining the counts for "Minor injury suspected," "Property damage only," and "Serious injury suspected" offers valuable insights into the distribution of severity across various contexts. The increased occurrence of serious injury suspected within topics such as "Roadway Crashes," "General Crashes," and "Traffic Violations" can be attributed to various factors. These topics encompass a wide array of roadway incidents inherently associated with a higher risk of serious injuries due to their diverse nature and characteristics. Roadway crashes often involve high-speed collisions, multi-vehicle accidents, or instances of reckless driving, all of which significantly elevate the likelihood of severe outcomes. Similarly, "General Crashes" may encompass rear-end collisions, side-impact crashes, or collisions at intersections, all with the potential for serious injuries depending on variables such as speed, vehicle type, and impact angle. Additionally, "Traffic Violations" may include infractions such as speeding, running red lights, or failure to yield, all of which are known to increase the risk of accidents and serious injuries. Moreover, these topics often involve interactions between various road users, including pedestrians, cyclists, and motor vehicle occupants, further intensifying the risk of serious injuries.

Topic	**KAB**	**C**	**O**
Roadway Crashes	18	43	6
General Crashes	5	56	9
Traffic Violations	5	14	3
Driveway Crashes	4	34	6
Crosswalk Crashes	3	57	13
Pedestrian Crashes	3	4	1
Cyclist Injuries	2	10	4
Vehicle Crashes with Bicycles	1	5	1

TABLE 3: Crash Severity by Topic

The prevalence of serious injuries within topics like "Roadway Crashes," "General Crashes," and "Traffic Violations" underscores the significant risk associated with these incidents. These topics cover a wide range of roadway collisions and violations, contributing to their high frequency due to their broad categorization. However, this broad classification poses challenges, particularly regarding the underreporting of bicycle-related crashes. Bicycle-related incidents often involve unique circumstances not fully captured within generalized topics like "Roadway Crashes." Factors such as misclassification, lack of standardized reporting, and minimal enforcement contribute to underreporting. Enhancing reporting protocols and implementing mechanisms for capturing bicycle-related incidents are crucial for accurately assessing their scope and implementing effective safety measures.

Interestingly, "Crosswalk Crashes" represent the highest number of minor injuries across this topics. This trend can be explained by several factors unique to collisions that occur within designated pedestrian and cyclist crossing areas. Firstly, these collisions often happen at lower speeds compared to incidents on open roadways, as drivers are typically more cautious in areas with pedestrian crossings. Despite the presence of traffic control devices like traffic signals or pedestrian signs, minor injuries can still occur due to factors such as driver inattention or pedestrian misjudgment of crossing times. Moreover, the layout and design of crosswalks, especially those located at complex intersections or with limited visibility, can further elevate the risk of minor injuries. Improving safety measures within crosswalks, such as enhancing signage, increasing visibility through better lighting, and enforcing traffic laws more rigorously, could effectively mitigate the risks associated with crosswalk collisions and reduce the occurrence of minor injuries. Overall, this relative comparison enables a nuanced understanding of the severity landscape across different SUV-bicycle incident topics, providing insights that can inform targeted safety measures and further investigative analyses within specific thematic areas.

Figure 4 provides a comprehensive analysis of bicycle crash statistics spanning five years (2017-2021), revealing distinct trends in the relative distribution of topics over time. Analyzing the temporal trends of bicycle-related incidents provides valuable insights into the frequency and dynamics of specific topics over time. Among the observed trends, certain topics exhibit notable fluctuations in frequency over the years, reflecting changes in patterns of bicycle-related incidents. For instance, "Crosswalk Crashes" show a fluctuating trend, with an increase in 2018 followed by a gradual decline in subsequent years. This pattern may suggest alterations in pedestrian and cyclist behaviors or modifications in infrastructure design, impacting the occurrence of incidents at crosswalks. Similarly, "Driveway Crashes" and "General Crashes" demonstrate varying frequencies over the years, with peaks observed in specific years. These fluctuations may be indicative of changes in traffic flow, driver behaviors, or environmental factors influencing crash occurrences. These changes can also be attributed the broad nature of these topics which describe a wide array of crashes. Additionally, "Traffic Violations" exhibit a pattern of varying frequencies, with peaks in certain years, indicating potential shifts in enforcement practices or compliance levels over time.

The average frequencies for each year reveal notable trends in bicycle-related incidents over the five-year period. The relatively lower average frequency in 2017 suggests a period of stability or potential underreporting, while the substantial increase in 2018 indicates a significant surge in incidents. Despite a slight decrease in 2019, incident frequencies remain elevated compared to previous years, with 2020 and 2021 showing persistent concerns for bicycle safety. These trends underscore the ongoing need for proactive measures to address safety issues and mitigate risks associated with bicycle-related incidents. Analyzing topics over time and observing the general shift in frequency of bicycle accidents offers valuable insights into the evolving landscape of safety risks. Identifying specific trends, such as fluctuations in the occurrence of "Bicycle Strikes on Roads" or spikes in "Road Travel Fails," enables a targeted understanding of critical areas for preventive measures. This temporal analysis allows policymakers and safety advocates to adapt strategies to address dynamic patterns and potential vulnerabilities, enhancing the effectiveness of preventive measures.

Figure 5 highlights possible similarities between each topic labeled by the BERTopic & LLM process. Identifying possible similarities in topics within a dataset of SUV-bicycle crash reports can be instrumental in gaining insights into severity patterns. A similarity matrix allows for a systematic examination of relationships between different topics, unveiling potential patterns or clusters that might not be immediately apparent. This analysis can help discern commonalities in contributing factors, circumstances, or locations across various crash scenarios. By understanding these similarities, analysts can identify overarching themes and recurring patterns, providing a more nuanced understanding of the factors influencing crash severity. This, in turn, enables targeted interventions, such as improved infrastructure design, awareness campaigns, or policy changes, to address specific patterns revealed by the similarity matrix and ultimately enhance safety for cyclists involved in SUV-related crashes. Here we can see that the trends in similarity from Figure 3

The similarities in topics are derived from the TF-IDF analysis, which highlights words that are both significant and common across different crash narratives in the dataset. In the context

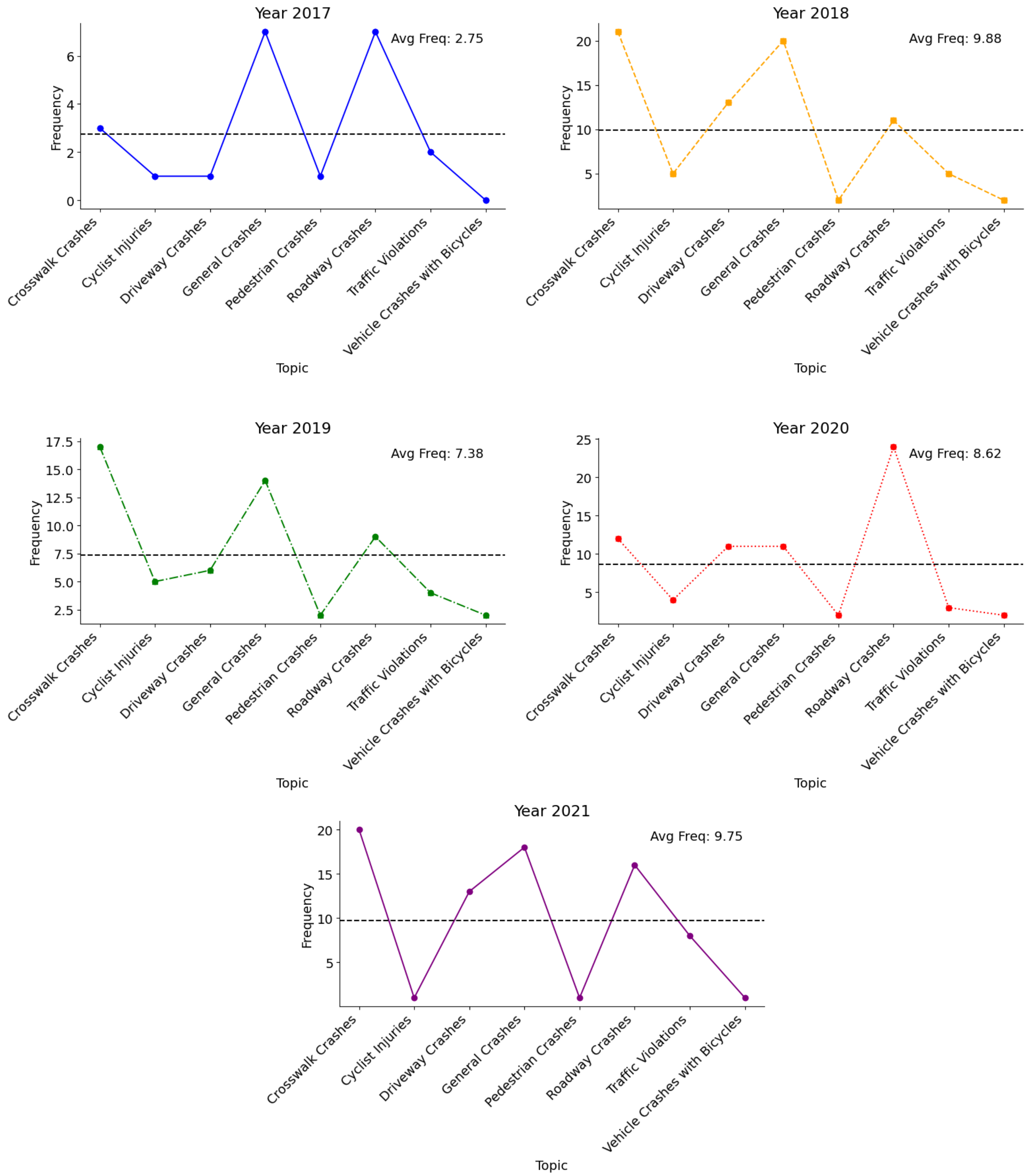

FIGURE 4: Temporal Frequency of Topics

of analyzing SUV-bicycle crash reports, significant and similar TF-IDF words play a crucial role. When topics share similar TF-IDF words, it indicates commonality in the language used to describe specific aspects of crashes. Given that the topics are derived from narratives within the focus area of crash reports, this similarity is expected and relevant. It suggests that certain words or phrases consistently appear in descriptions of SUV-bicycle crashes, forming thematic clusters or

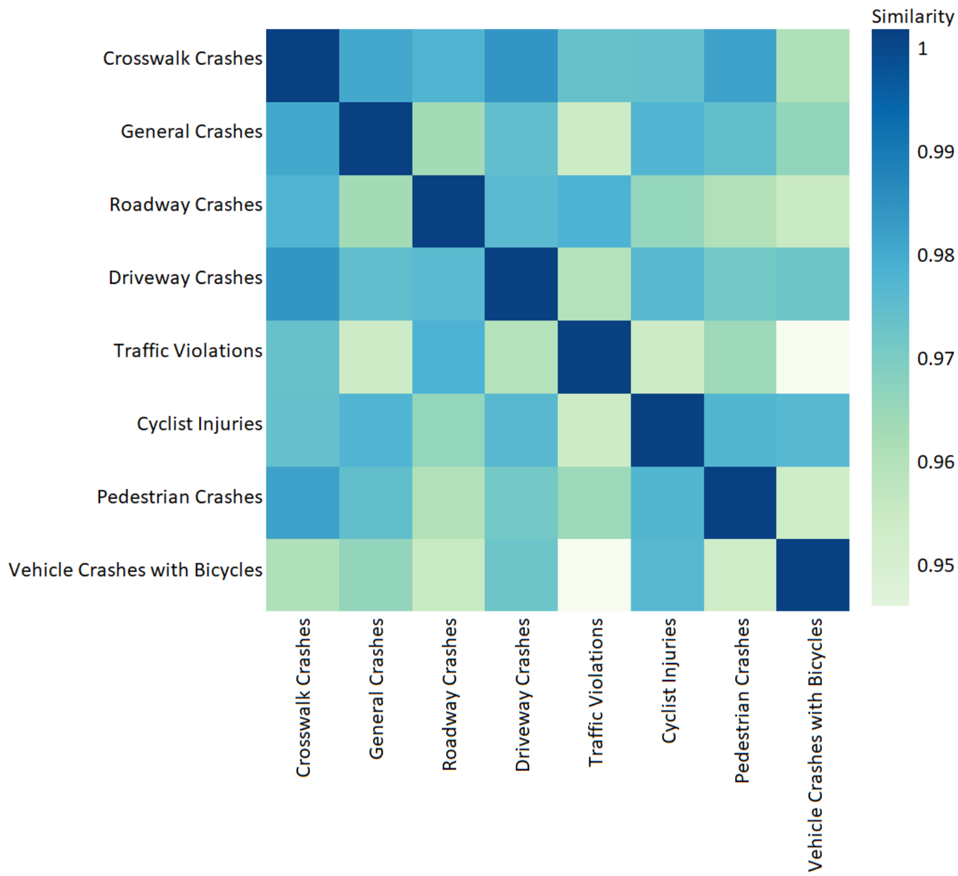

FIGURE 5: Topic Similarity

patterns. Recognizing these similarities provides a basis for grouping topics with shared linguistic characteristics, allowing for a greater understanding of common factors influencing crash severity. It helps to unveil underlying themes and patterns within the dataset, aiding in the identification of key contributing factors and the development of targeted interventions to enhance cyclist safety in SUV-related crashes. Additionally, this matrix also provides some insights into some of the limitations of the topic modeling, where some topics that should be similar, such as those involving intersections, were not marked as having a high similarity.

CONCLUSION

This study addresses the critical concern of transportation safety, specifically focusing on SUV-bicycle crashes. We used the power of advanced models like BERTopic and LLMs, to uncover thematic clusters. The study successfully revealed thematic clusters within the dataset, shedding light on the circumstances surrounding SUV-bicycle incidents. By employing topic modeling techniques, it identified common themes and patterns associated with different severity levels, ranging from minor injuries to serious injuries. The integration of LLM labels and advanced clustering methods allowed for a detailed exploration of the content, providing a comprehensive understanding of the dataset. The findings from this study have implications, especially in informing and guiding policy formulation and safety interventions. The ability to pinpoint common themes and severity patterns associated with SUV-bicycle crashes is invaluable. It can equip policymakers and safety advocates with empirical evidence to craft targeted strategies aimed at mitigating risks and safeguarding cyclists. By understanding how and when specific crash scenarios are more likely to occur, interventions can be timely and more effectively aligned with the evolving landscape of road use.

Despite the advancements and insights provided by this study, it's important to acknowledge its limitations and the avenues it opens for future research. One of the primary constraints lies in the reliance on reported crash data, which may not capture all incidents, particularly those that go unreported. Additionally, the NLP techniques, while powerful, depend heavily on the quality and completeness of the data available. This highlights the need for continuous improvement in data collection and reporting methods to ensure a more comprehensive analysis. Looking ahead, future studies could explore integrating more diverse data sources, including social media and eyewitness accounts, to enrich the dataset and provide a more nuanced understanding of SUV-bicycle crashes. Further refinement of NLP models to better capture the subtleties of language used in crash reports could also enhance the accuracy of thematic analysis. Moreover, an exciting direction for future work involves the development of forecasting models to predict the occurrence of specific crash topics across different seasons or months. Such predictive analytics could offer preemptive insights, enabling stakeholders to anticipate and mitigate potential risks with seasonally tailored safety measures. Ultimately, this research paves the way for a more detailed exploration of factors contributing to transportation safety, encouraging the development of innovative solutions that leverage the latest advancements in technology and data analysis to protect vulnerable road users.

REFERENCES

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the surprising behavior of distance metrics in high dimensional space, in: Van den Bussche, J., Vianu, V. (Eds.), Database Theory — ICDT 2001, Springer. pp. 420–434. doi:10.1007/3-540-44503-X_27.
2. Amoros, E., Martin, J.L., Laumon, B., 2006. Under-reporting of road crash casualties in france 38, 627–635. URL: https://www.sciencedirect.com/science/article/pii/S0001457505001910, doi:10.1016/j.aap.2005.11.006.
3. Aultman-Hall, L., Kaltenecker, M.G., 1999. Toronto bicycle commuter safety rates 31, 675–686. URL: https://www.sciencedirect.com/science/article/pii/S0001457599000287, doi:10.1016/S0001-4575(99)00028-7.
4. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful?, in: Beeri, C., Buneman, P. (Eds.), Database Theory — ICDT'99, Springer. pp. 217–235. doi:10.1007/3-540-49257-7_15.
5. Blei, D.M., 2003. Latent dirichlet allocation .
6. Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates, in: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining, Springer. pp. 160–172. doi:10.1007/978-3-642-37456-2_14.
7. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., Blei, D., 2009. Reading tea leaves: How humans interpret topic models, in: Advances in Neural Information Processing

Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html.

8. Das, S., Oliaee, A.H., Le, M., Pratt, M.P., Wu, J., 2023. Classifying pedestrian maneuver types using the advanced language model 2677, pp 599–611. URL: https://trid.trb.org/view/2134886, doi:10.1177/03611981231155187. publisher: Sage Publications, Incorporated.

9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. URL: http://arxiv.org/abs/1810.04805, doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].

10. Doogan, C., Buntine, W., 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures, in: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics. pp. 3824–3848. URL: https://aclanthology.org/2021.naacl-main.300, doi:10.18653/v1/2021.naacl-main.300.

11. Drosouli, I., Voulodimos, A., Mastorocostas, P., Miaoulis, G., Ghazanfarpour, D., 2023. TMD-BERT: A transformer-based model for transportation mode detection 12, 581. URL: https://www.mdpi.com/2079-9292/12/3/581, doi:10.3390/electronics12030581. number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

12. Elvik, R., Mysen, A., 1999. Incomplete accident reporting: Meta-analysis of studies made in 13 countries 1665, 133–140. URL: https://doi.org/10.3141/1665-18, doi:10.3141/1665-18. publisher: SAGE Publications Inc.

13. Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press. pp. 226–231.

14. Ghasemi, N., Acerra, E., Vignali, V., Lantieri, C., Simone, A., Imine, H., 2020. Road safety review update by using innovative technologies to investigate driver behaviour 45, 368–375. URL: https://www.sciencedirect.com/science/article/pii/S2352146520301903, doi:10.1016/j.trpro.2020.03.028.

15. Gildea, K., Hall, D., Simms, C., 2021. Configurations of underreported cyclist-motorised vehicle and single cyclist collisions: Analysis of a self-reported survey 159, 106264. URL: https://www.sciencedirect.com/science/article/pii/S0001457521002955, doi:10.1016/j.aap.2021.106264.

16. Grootendorst, M., 2024. MaartenGr/BERTopic. URL: http://arxiv.org/abs/2203.05794. original-date: 2020-09-22T14:19:29Z.

17. Joachims, T., 1996. A probabilistic AnalysisToexf tthCeaRteogcocrhizioatAiolngorithm with TFIDF for URL: https://apps.dtic.mil/sti/citations/ADA307731.

18. Lee, S., Arvin, R., Khattak, A.J., 2023. Advancing investigation of automated vehicle crashes using text analytics of crash narratives and bayesian analysis 181, 106932. URL: http://www.sciencedirect.com/science/article/pii/S0001457522003670, doi:10.1016/j.aap.2022.106932. publisher: Elsevier.

19. McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: Uniform manifold approx-

imation and projection 3, 861. URL: https://joss.theoj.org/papers/10.21105/joss.00861, doi:10.21105/joss.00861.

20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. URL: http://arxiv.org/abs/1310.4546, doi:10.48550/arXiv.1310.4546, arXiv:1310.4546 [cs, stat].

21. Ohio Department of Transportation, . Ohio department of transportation | ohio.gov. URL: https://www.transportation.ohio.gov/.

22. Oliaee, A.H., Das, S., Liu, J., Rahman, M.A., 2023. Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types 3, 100007. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000043, doi:10.1016/j.nlp.2023.100007.

23. Pandove, D., Goel, S., Rani, R., 2018. Systematic review of clustering high-dimensional and large datasets 12, 16:1–16:68. URL: https://doi.org/10.1145/3132088, doi:10.1145/3132088.

24. Pealat, C., Bouleux, G., Cheutet, V., 2021. Improved time-series clustering with UMAP dimension reduction method, in: ICPR 2020 - 25th Internation conference in Pattern Recognition. URL: https://hal.science/hal-03188503.

25. 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), Association for Computational Linguistics. pp. 1532–1543. URL: https://aclanthology.org/D14-1162, doi:10.3115/v1/D14-1162.

26. Qin et al, 2021. Using text data from the DT4000 to enhance crash analysis .

27. Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics. pp. 3982–3992. URL: https://aclanthology.org/D19-1410, doi:10.18653/v1/D19-1410.

28. Reynolds, C.C., Harris, M.A., Teschke, K., Cripton, P.A., Winters, M., 2009. The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature 8, 47. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2776010/, doi:10.1186/1476-069X-8-47.

29. Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis 20, 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257, doi:10.1016/0377-0427(87)90125-7.

30. Shinar, D., Valero-Mora, P., van Strijp-Houtenbos, M., Haworth, N., Schramm, A., De Bruyne, G., Cavallo, V., Chliaoutakis, J., Dias, J., Ferraro, O.E., Fyhri, A., Sajatovic, A.H., Kuklane, K., Ledesma, R., Mascarell, O., Morandi, A., Muser, M., Otte, D., Papadakaki, M., Sanmartín, J., Dulf, D., Saplioglu, M., Tzamalouka, G., 2018. Underreporting bicycle accidents to police in the COST TU1101 international survey: Cross-country comparisons and associated factors 110, 177–186. doi:10.1016/j.aap.2017.09.018.

31. Tarr, D., 2018. dominictarr/random-name. URL: https://github.com/dominictarr/random-name. original-date: 2012-09-24T00:05:04Z.

32. Taylor, W.L., 1953. "cloze procedure": A new tool for measuring readability 30, 415–433. URL: http://journals.sagepub.com/doi/10.1177/107769905303000401, doi:10.1177/107769905303000401.

33. Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I., 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. URL: http://arxiv.org/abs/2010.08240, doi:10.48550/arXiv.2010.08240, arXiv:2010.08240 [cs].

34. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. URL: http://arxiv.org/abs/2307.09288, doi:10.48550/arXiv.2307.09288, arXiv:2307.09288 [cs].

35. 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 34, 22–30. URL: https://www.sciencedirect.com/science/article/pii/S0386111210000105, doi:10.1016/j.iatssr.2010.07.002.

36. Valcamonico, D., Baraldi, P., Amigoni, F., Zio, E., 2022. A framework based on natural language processing and machine learning for the classification of the severity of road accidents from reports , 1748006X221140196URL: https://doi.org/10.1177/1748006X221140196, doi:10.1177/1748006X221140196. publisher: SAGE Publications.

37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. URL: http://arxiv.org/abs/1706.03762, doi:10.48550/arXiv.1706.03762, arXiv:1706.03762 [cs].

38. Wali, B., Khattak, A.J., Ahmad, N., 2021. Injury severity analysis of pedestrian and bicyclist trespassing crashes at non-crossings: A hybrid predictive text analytics and heterogeneity-based statistical modeling approach 150, 105835. URL: http://www.sciencedirect.com/science/article/pii/S0001457520316559, doi:10.1016/j.aap.2020.105835. publisher: Elsevier.

39. Weng, Y., Das, S., Paal, S.G., 2023. Applying few-shot learning in classifying pedestrian crash typing 2677, pp 563–572. URL: https://trid.trb.org/view/2138544, doi:10.1177/03611981231157393. publisher: Sage Publications, Incorporated.

40. World Health Organization, 2023. Global Status Report on Road Safety: Time for Action. World Health Organization. URL: Snapshot:C\protect\protect\leavevmode@ifvmode\kern+.2222em\relax\Users\Jett\Zotero\storage\AVVAQDC4\road-traffic-injuries.html:text/html. google-Books-ID: Ndrf6DuCQHMC.

41. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., 2023. C-pack: Packaged resources

to advance general chinese embedding. URL: http://arxiv.org/abs/2309.07597, arXiv:2309.07597 [cs].

42. Yang, X., Gao, J., Xue, W., Alexandersson, E., 2023. PLLaMa: An open-source large language model for plant science. URL: https://arxiv.org/abs/2401.01600v1.