Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Analyzing relationships between latent topics in autonomous vehicle crash narratives and crash severity using natural language processing techniques and explainable XGBoost

Pei Li[a], Sikai Chen[a,*], Lishengsa Yue[b], Yuan Xu[a], David A. Noyce[a]

[a] *Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI 53706, United States of America*
[b] *The Key Laboratory of Road and Traffic Engineering, Ministry of Education Tongji University, Shanghai, China*

ARTICLE INFO

Keywords:
Autonomous vehicle safety
Machine learning
Natural language processing
Topic modeling

ABSTRACT

Safety is one of the most essential considerations when evaluating the performance of autonomous vehicles (AVs). Real-world AV data, including trajectory, detection, and crash data, are becoming increasingly popular as they provide possibilities for a realistic evaluation of AVs' performance. While substantial research was conducted to estimate general crash patterns utilizing structured AV crash data, a comprehensive exploration of AV crash narratives remains limited. These narratives contain latent information about AV crashes that can further the understanding of AV safety. Therefore, this study utilizes the Structural Topic Model (STM), a natural language processing technique, to extract latent topics from unstructured AV crash narratives while incorporating crash metadata (i.e., the severity and year of crashes). In total, 15 topics are identified and are further divided into behavior-related, party-related, location-related, and general topics. Using these topics, AV crashes can be systematically described and clustered. Results from the STM suggest that AVs' abilities to interact with vulnerable road users (VRUs) and react to lane-change behavior need to be further improved. Moreover, an XGBoost model is developed to investigate the relationships between the topics and crash severity. The model significantly outperforms existing studies in terms of accuracy, suggesting that the extracted topics are closely related to crash severity. Results from interpreting the model indicate that topics containing information about crash severity and VRUs have significant impacts on the model's output, which are suggested to be included in future AV crash reporting.

1. Introduction

Autonomous vehicles (AVs) are capable of perceiving the surrounding environments and making decisions with little or no human input (Hu et al., 2020). They are believed to have the potential to significantly improve traffic safety and efficiency by eliminating the possibility of human errors, including distraction, impairment, and fatigue (NHTSA, 2020; NSTC, USDOT, 2020). Over 1,400 AVs have been in the process of testing by more than 80 companies in the US as of 2019 (Etherington, 2019). Among all the factors, safety is the most crucial consideration while designing AVs, and various Advanced Driver Assistance Systems (ADAS) have been designed to make safer AVs.

Despite the efforts to make safe AVs, they are still involved in crashes in real-world driving conditions. Understanding the characteristics of these crashes becomes a crucial task for identifying critical issues in AVs and providing suggestions for AV manufacturers. The California Department of Motor Vehicles (DMV) has made it mandatory for AV manufacturers to report any crashes that resulted in property damage, bodily injury, or death within 10 days of crashes since 2014 (California DMV, 2022). California DMV defines AVs as vehicles that can drive without active physical control or monitoring by a human operator. This definition corresponds to the automation levels 3 to 5 defined by SAE (International, 2021), representing automation features that can drive vehicles under limited or all conditions and require little or no human intervention. The AV crash data is publicly available through the website of California DMV. A crash report provides structured information including the date, time, severity, and type of the crash. In addition, it provides a detailed description of the crash, such as pre-crash behavior, location of the crash, road users involved in the crash, etc. Since the release of this data, extensive studies have been conducted, focusing on the exploratory analysis of AV crashes using

[*] Corresponding author.
E-mail addresses: pei.li@wisc.edu (P. Li), sikai.chen@wisc.edu (S. Chen), 2014yuelishengsa@tongji.edu.cn (L. Yue), xu353@wisc.edu (Y. Xu), danoyce@wisc.edu (D.A. Noyce).

https://doi.org/10.1016/j.aap.2024.107605
Received 10 October 2023; Received in revised form 3 March 2024; Accepted 27 April 2024
Available online 13 May 2024

structured crash data (Favarò et al., 2017; Teoh and Kidd, 2017; Leilabadi and Schmidt, 2019; Wang and Li, 2019; Lv et al., 2017). For example, rear-end crashes were suggested as the most common crashes, most AV crashes caused no or minor injuries, etc. The unstructured crash narratives, on the contrary, have not been fully investigated and require further analysis as they may contain latent information about AV crashes. Such information is expected to facilitate the understanding of AV crashes and provide suggestions for improving AV safety. Moreover, the latent information may be associated with crash severity. Interpreting these relationships would be beneficial for comprehending contributing factors to AV crash severity. Therefore, this study employs a topic modeling method to discover the latent topics from unstructured AV crash narratives. Topic modeling is a Natural Language Processing (NLP) technique that refers to a generative model of word counts and allows for rich latent topics to be automatically inferred from the unstructured text data (Roberts et al., 2014). In the framework of topic modeling, a topic is defined as a mixture of words where each word has a probability of belonging to a topic. A document (i.e., crash narrative) can be defined as a mixture of topics. For each document, a data-generating process is defined, then the data will be used to find the most likely values for the parameters within the model. Topic models can extend the understanding of text over pre-defined fields, which makes them ideal choices for exploring novel information. Topic modeling has been widely applied to various aspects, such as political analysis, consumer analysis, activity pattern classification, etc. (Grimmer, 2010; Ghazizadeh et al., 2014; Hasan and Ukkusuri, 2014). In addition, several studies (Roque et al., 2019; Brown, 2015; Kwayu et al., 2021; Alambeigi et al., 2020) have utilized topic modeling methods for traffic safety analysis to identify latent topics from crash narratives. Latent Dirichlet Allocation (LDA) is the most commonly used topic modeling model among existing studies. However, one limitation of LDA is that it cannot incorporate metadata (e.g., the year and severity of crashes) into the topic distributions assigned to crash narratives. Therefore, this study utilizes the Structural Topic Model (STM), an extension of LDA, to incorporate metadata while modeling crash narratives (Roberts et al., 2013), which aims to make the identified topics better reflect the temporal and severity information about crashes.

In summary, most existing studies have been focusing on general descriptive statistics about AV crashes using structured crash data. The unstructured crash narratives provide detailed information about AV crashes but have not been effectively utilized in existing studies as they are difficult to prepare and interpret. Results from similar studies in NLP indicate that topic modeling is an effective method for extracting latent information from text data. Inspired by them, this study utilizes a topic modeling technique, STM, to identify the latent topics from AV crash narratives. Topic modeling facilitates the understanding of AV crashes by converting unstructured crash narratives to structured topics, making it easier to manage and navigate through information, reducing the dimensionality of crash narratives, and capturing essential information while discarding less important details. Moreover, an explainable machine learning model is developed to further investigate the relationships between topics and crash severity, aiming to understand the contributions of topics to crash severity and provide suggestions to improve AV safety.

2. Literature review

2.1. AV crash analysis

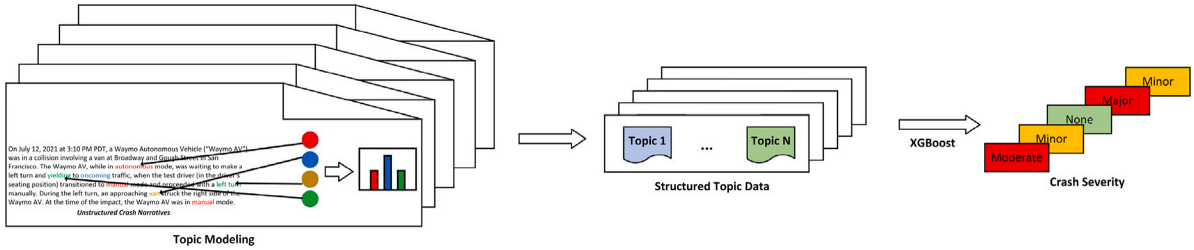
Existing studies have analyzed the patterns of AV crashes and the contributing factors to them (Favarò et al., 2017; Teoh and Kidd, 2017; Leilabadi and Schmidt, 2019; Wang and Li, 2019; Lv et al., 2017). Specifically, Favarò et al. (2017) analyzed AV crash and disengagement data from 2014 to 2017 in California. Results suggested that rear-end crashes were the most common among all crashes. Moreover, the limitation of detection was suggested as a major contributing factor to AV crashes. Teoh and Kidd (2017) analyzed Google AV crash data during 2009–2015 and suggested that AVs had a lower rate of crashes per million vehicle miles traveled (i.e., 2.19) than Human-driven Vehicle (HDVs) (i.e., 6.06). In addition, AVs could perform more safely than HDVs in certain conditions but would continue to be involved in crashes with HDVs due to various reasons. Leilabadi and Schmidt (2019) analyzed AV crash and disengagement data from 2014 to 2018 in California. Results suggested that most AV crashes only resulted in minor damage. In addition, the quality of the road's surface could manipulate and confound AVs thus resulting in crashes. Wang and Li (2019) investigated the contributing factors from perception, planning, and control stages to AV disengagements. Results from logistic regression and classification tree models suggested that the planning issue (i.e., undesired behaviors from road users, computation issues of planning, and completing a lane change) was the most significant factor that affected AV disengagements. Moreover, the lack of radar and LiDAR sensors also had negative impacts on AV disengagements. In addition, several studies have investigated the relationships between contributing factors and the frequency or severity of AV crashes. For example, Boggs et al. (2020) developed a Bayesian logistic model to analyze contributing factors to rear-end AV crashes. Results indicated several factors were positively correlated with the likelihood of rear-end AV crashes, such as the engagement of autonomous driving systems, operating AVs on mixed land-use environments, etc. Zhu and Meng (2022) utilized a decision tree model for identifying crash severity using AV crash data from 2018 to 2021 in California. Basic information about AV crashes, including time, location, weather conditions, and lighting conditions was used for developing the model. Crash severity was coded as a binary variable where injury and non-injury crashes were classified using the model. On average, the model achieved a sensitivity of 0.69, a specificity of 0.65, and an accuracy of 0.66.

In summary, most existing studies have been focusing on general statistics about AV crashes using formatted AV crash data. The unstructured crash narratives have not been effectively utilized as they are difficult to prepare and interpret, which requires further investigation since such data may provide undiscovered characteristics about AV crashes. In addition, although some studies attempted to identify contributing factors to AV crash severity, the accuracy needs further improvement as only basic attributes about AV crashes were used. Moreover, crash severity should be further classified into multi-class rather than binary class to better understand the relationships between contributing factors and crash severity.

2.2. Topic modeling

Topic modeling is an NLP task that aims to identify latent topics from text data using statistical models such as LDA, Latent Semantic Analysis (LSA), STM, etc. Moreover, the recent advances in Large Language Models (LLM), represented by models including Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPTs), have provided new ways of extracting topics from texts.

Although both statistical and LLM models can be used for topic modeling, the differences between them are significant and should be carefully examined. In particular, LLMs rely on pre-trained models which are developed based on massive amounts of data. Therefore, LLMs perform well in general topic modeling applications but may fail in domain-specific areas unless re-trained using substantial data from those domains. Differently, statistical models need to be trained on the specific dataset, and therefore, present patterns unique to that context. Given the study's focus on AV crash narratives, which are both limited in number and highly domain-specific, statistical models are the more appropriate choice for discovering topics. Models like LDA and STM are well-suited for handling smaller datasets and capturing domain-specific nuances, aligning effectively with the study's objectives. In addition,



On July 12, 2021 at 3:10 PM PDT, a Waymo Autonomous Vehicle ("Waymo AV") was in a collision involving a van at Broadway and Gough Street in San Francisco. The Waymo AV, while in autonomous mode, was waiting to make a left turn and yielding to oncoming traffic, when the test driver (in the driver's seating position) transitioned to manual mode and proceeded with a left turn manually. During the left turn, an approaching van struck the right side of the Waymo AV. At the time of the impact, the Waymo AV was in manual mode.

Fig. 1. Research diagram of this study, the unstructured crash narratives are converted to collections of topics using topic modeling, and an XGBoost model is developed to estimate the relationships between topics and crash severity.

LDA and STM were widely used in topic modeling for HDV crashes. For example, Roque et al. (2019) used LDA to identify patterns related to run-off-road crashes in Irish using reports from Road Safety Inspection (RSI). Specifically, an RSI report contains a brief description of each safety issue, an informal risk assessment of it, and recommendations to deal with it. The authors divided documents into problem and solution sets and performed LDA modeling in each set. The experimental results suggested that the frequency of topics related to roadside safety was higher in the problem set than in the solution set, indicating that it may be easier to identify roadside safety-related problems than to provide solutions. Brown (2015) used LDA to understand contributors to rail crashes. By using topics extracted from LDA, the accuracy of crash cost prediction was significantly improved, indicating that the extracted topics could reflect contributors to crashes. Kwayu et al. (2021) utilized STM to analyze narratives of fatal crashes that occurred in Michigan from 2009 to 2018. In total, twenty-five topics were discovered using STM from 9202 crash narratives. In addition, these topics could be categorized into three groups, event-based, location-based, and involved parties-based topics. The authors suggested that the generated topics could be utilized to classify crash types, including rear-end, sideswipe, and head-on with an accuracy of 99.2%, indicating that the extracted topics were closely related to crash types. Alambeigi et al. (2020) used a probabilistic topic modeling approach to analyze narratives of AV crashes. Results suggested that AV crashes could be classified into several categories such as driver-initiated transition crashes, sideswipe crashes during left-side overtakes, etc.

In summary, existing studies have suggested that topic modeling is an effective way of discovering latent information about crashes from crash narratives. This latent information could be used to better understand contributing factors to crashes. In addition, compared to LDA, STM is preferred for modeling crash narratives due to its ability to incorporate crash metadata including the time, severity, and type of crashes.

3. Methods

Fig. 1 presents the research diagram of this study. The unstructured crash narratives are first prepared by removing punctuations, lowercasing, and tokenizing. They are then transformed into structured topics using STM. Specifically, each crash narrative is converted into a set of topics with the sum of their proportions as one. Lastly, an XGBoost model is applied to estimate the relationships between the topics and crash severity. Detailed methods are presented in the following sections.

3.1. Structural topic model

Topic models are machine learning models for discovering latent information (i.e., topics) from text data. To incorporate crash metadata into the model, this study utilizes STM for modeling crash narratives. STM is a combination and extension of three existing topic models, the correlated topic model (CTM) (Blei and Lafferty, 2007), the Dirichlet-Multinomial Regression (DMR) model (Mimno and McCallum, 2008), and the Sparse Additive Generative (SAGE) model (Eisenstein et al., 2011). The diagram of STM is shown in Fig. 2, where X is the topic prevalence covariates, θ is the narrative-topic proportion, Z is the per-word topic assignment, w is the word, β is the topic-word distribution, and Y is the topic content covariates. STM has three essential components, a topic prevalence model which controls how words are allocated to topics, a topic content model which controls the frequency of words in topics, and a language model which combines the prevalence and content models to produce the actual words (Roberts et al., 2016). Given a crash narrative d that has a vocabulary of size V, an STM with K topics can be summarized as (Roberts et al., 2016, 2013):

- The topic proportion θ_d is drawn from a logistic-normal generalized linear model based on a 1 by p vector of topic prevalence covariates X_d, where $\sum$ is a $K-1$ by $K-1$ covariance matrix and γ is a p by $K-1$ matrix of coefficients, as shown in Eq. (1).

$$\begin{gathered}\theta_d \sim LogisticNormal(\mu_d, \sum)\\ \mu_{d,k} = X_d\gamma_k\end{gathered} \tag{1}$$

- The topic-word distribution β in each topic is a random variable conditional on baseline word distribution m_v, topic k, and covariates y as shown in Eq. (2), where m_v is the estimated log-transformed rate of occurrence of terms v in the document, k^k, k^{y_d}, and $k^{y_d,k}$ are the topic-specific deviation, the covariate group deviation and the interaction between the topic deviation and covariate deviation, respectively.

$$\beta_{d,v}{}^k \propto exp(m_v + k^k + k^{y_d} + k^{y_d,k}) \tag{2}$$

- For each word in the narrative ($n \in 1,\dots,N_d$), the word's topic assignment $z_{d,n}$ is based on the narrative-specific distribution over topics θ_d, while the word assignment $w_{d,n}$ is conditional on the topic chosen, as shown in Eq. (3).

$$\begin{gathered}z_{d,n} \sim Multinomial(\theta_d)\\ w_{d,n} \sim Multinomial(\beta_{d,k=z_{d,n}})\end{gathered} \tag{3}$$

The main advantage of STM over other topic models is it incorporates metadata and discovers relationships between topics and metadata. The severity and year of crashes are used as metadata in this study as severity is one of the most important considerations in crash analysis. The year of crashes is used to allow the model to be aware of the temporal information. In addition, a spline transformation is applied to the year of the crash to allow a non-linear effect of this variable.

3.1.1. Selecting the number of topics

STM is an unsupervised learning model and the number of topics it contains is subjective to the data it is trained. To determine the number of topics in an STM model, this study utilizes several metrics including the residual, exclusivity, and semantic coherence of the model. Specifically, the residual of an STM represents the estimated sample dispersion (Taddy, 2012). The exclusivity and semantic coherence are estimated in terms of each topic in an STM. For one topic,

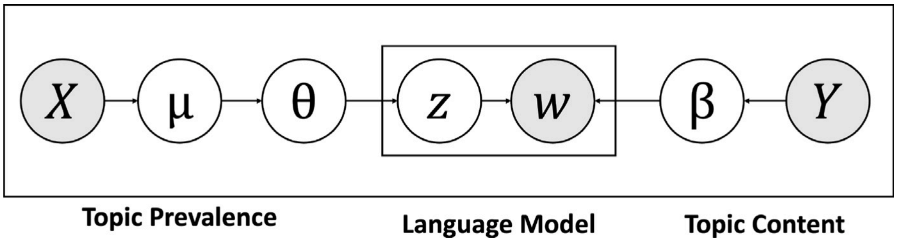


Fig. 2. The diagram of STM (Roberts et al., 2013).

exclusivity indicates the exclusivity of words in a given topic (Bischof and Airoldi, 2012). Semantic coherence is maximized when the most probable words in a given topic frequently co-occur together in a document (Mimno et al., 2011). In general, an optimal STM should have a small value of residual, a large value of exclusivity, as well as a large value of semantic coherence.

3.1.2. *Interpreting topics*

Topics are combinations of words, therefore, the most straightforward way to interpret a topic is to understand the words that it consists of. However, a topic may contain many words while only part of them could represent the meaning of it. Therefore, several metrics, including probability, FREX, lift, and score, are commonly used to assign weights to words in a topic, and words that have higher weights are used to interpret the topic. Specifically, the probability of a word in a topic is estimated based on the topic-word distribution, which is represented by the parameter β (Eq. (2)). For example, β_i represents the word distribution of topic i, which has N elements (i.e., the number of words) and sums up to 1, and $\beta_{k,v}$ represents the probability of seeing word v conditional on topic k. FREX (Frequency-Exclusivity) (Bischof and Airoldi, 2012) weights words by their overall frequency and how exclusive they are to the topic. FREX attempts to find words that are both frequent and exclusive to a topic of interest. Eq. (4) shows the estimation of a word's FREX, where F represents the frequency of the word, E indicates the exclusivity of the word, and w is the weight which is set to 0.7 to favor exclusivity.

$$FREX = (\frac{w}{F} + \frac{1-w}{E})^{-1} \tag{4}$$

Lift (Taddy, 2013) is calculated by dividing the topic-word distribution by the empirical word count probability distribution. The word count is a vector indicating the number of times each word appears in the corpus. Lift gives higher weight to words that appear less frequently in other topics. Similar to lift, score (Chang, 2011) divides the log frequency of the word in the topic by the log frequency of the word in other topics. Considering the similarity between lift and score, this study eventually uses probability, FREX, and lift as metrics to interpret topics.

3.2. *Crash severity analysis*

Crash severity analysis could be conducted as a multi-class classification problem, and accuracy and interoperability are two essential considerations when modeling crash severity. Therefore, this study utilizes XGBoost for this task as it has achieved high accuracy in similar studies (Shi et al., 2022; Li and Abdel-Aty, 2022) compared to traditional statistical models. In addition, XGBoost is a decision tree-based model that is much easier to explain than other "black-box" machine learning models. The explanation of the model can quantify and understand the relationships between the topics and crash severity, which is beneficial for providing suggestions for improving AV safety and reducing AV crash severity. Lastly, XGBoost requires much less computing resources compared to deep learning models such as convolutional neural networks. Specifically, XGBoost is proposed by Chen and Guestrin (2016) as a gradient boosting model. It introduces regularization, parallel processing, and other techniques to improve the performance of the original gradient boosting. The trees in an XGBoost model are built in a sequential way so that each new tree corrects the errors of the previous tree. For a given data, an XGBoost uses K additive functions to estimate $\hat{y}$ as shown in Eq. (5), where $\mathscr{F}$ is the space of regression trees. Eq. (6) gives an example of a tree in XGBoost, where q represents the structure of each tree and w represents the leaf weights. The objective of XGBoost is to minimize the value of the regularized objective function in Eq. (7), where l is a differentiable convex loss function that estimates the difference between y_i and $\hat{y}_i$, Ω is a regularization term that penalizes the model's complexity as shown in Eq. (8). The usage of the regularization function prevents the model from becoming over-complicated, where T is the number of leaves, λ, and γ are the regularization parameters.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathscr{F} \tag{5}$$

$$f_k(x) = w_{q(x)} \tag{6}$$

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k-1}^{K} \Omega(f_k) \tag{7}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{8}$$

To develop the XGBoost model in this study, data containing topics and crash severity are first randomly divided into training and test data five times. Moreover, the training data are balanced to overcome the data imbalance issue using the synthetic minority over-sampling technique (Chawla et al., 2002). Then, for each data split, an XGBoost model is trained on the training data with its hyperparameters optimized according to its performance on the test data as shown in Fig. 3. The log loss is used for optimizing the hyperparameters as $\sum_j^M(-\frac{1}{N}\sum_i^N y_{ij} \cdot Log(p_{ij}))$, where N is the number of samples, M is the number of classes, y_{ij} is the binary variable with the expected classes and p_{ij} is the estimated probability for the ith sample and the jth class. Several hyperparameters including the number of trees, learning rate, the maximum depth of a tree, etc. are optimized in the training. Lastly, the model's precision, recall, and F1-score on training and test data are estimated and averaged over the five splits.

4. Data

4.1. *Data preparation*

Data used in this study are obtained from California DMV (California DMV, 2022). The California DMV has been collecting AV crash data since 2014. As of June 2, 2023, the DMV has received 604 AV crash reports, which provide detailed information about AV crashes, including the location, time, severity, and type of crashes. In addition, each crash report contains a detailed narrative that provides additional information about it, including pre-crash events, road-user behavior, etc., as shown in Fig. 4. Crash reports are downloaded in PDF format. Then, an R package $pdftools$ (Ooms, 2022) is used to extract narratives

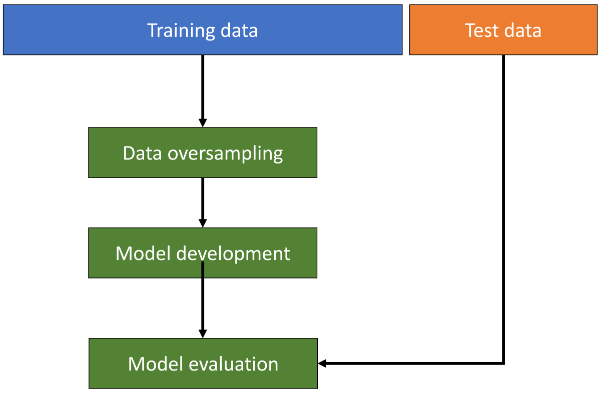


Fig. 3. The flowchart of develeoping the model.

Table 1
Words removed from the crash narratives.

Types	Words
Manufacturer	waymo, google, zoox, apple, lyft, argo, …
Location	los, altos, antonio, san, francisco, mountain, view, southeast, westbound, northeast, …
Report	pst, police, driver, call, information, mph, august, pdt, incident, report, day, driver, …

of crashes from crash reports, as well as the year, severity, and other information about crashes. The extracted data are prepared in a spreadsheet format where each crash has its identification, narrative, severity, and time. In addition, the crash narratives are prepared as shown in Fig. 5. First, stop words, punctuations, and numbers are removed from narratives as they usually do not provide meaningful and distinct information about individual crashes. Moreover, some other words that are commonly found among crash reports but do not provide practical information about crashes are also removed from crash narratives, including location-, manufacturer-, and report-related words, as shown in Table 1. After cleaning the narrative data, each crash narrative is converted into a bag-of-word matrix and a metadata matrix. A bag-of-words matrix is a numerical representation of the crash narrative that the STM model can understand and analyze as shown in Fig. 5. A metadata matrix contains temporal and severity information about the crash.

4.2. Data summary

This study has used AV crash data from 2018 to 2022 as they account for nearly 82% of the total AV crashes (California DMV, 2022). The format of crash reports was updated in 2017, thus data from 2014 to 2017 are not used in this study. Fig. 6 shows the yearly distribution of AV crashes. The number of AV crashes has notable temporal variations. In particular, 2020 has the lowest number of AV crashes, probably due to the impact of COVID-19. Therefore, it is necessary to incorporate the temporal information while modeling AV crash narratives. In total, 463 crashes are used in this study after removing crashes that have incomplete information. Specifically, 41 crashes did not cause injuries, 344 crashes caused minor injuries, and 78 crashes caused moderate or major injuries. After preparing crash narratives using the methods presented in previous sections, crash narrative data are converted into a vocabulary vector, a narrative matrix, and a metadata matrix. Specifically, a vocabulary vector containing the words associated with the word index that is generated for the entire narrative data, and the words are ordered alphabetically. Moreover, a narrative matrix containing all narratives is generated using the bag-of-word transformation. Lastly, a metadata matrix containing metadata including the year and severity of each crash is also created. Fig. 7 shows the word cloud of the vocabulary vector. Words such as street, rear, bumper, stop, minor, etc. are the most frequent words as expected, suggesting most crashes occur on public streets and rear-end crashes are the most common crashes. Moreover, words such as disengage, sensor, and manual have a relatively high frequency as these words are related to the perception and control components of AVs.

5. Results

5.1. Topic modeling

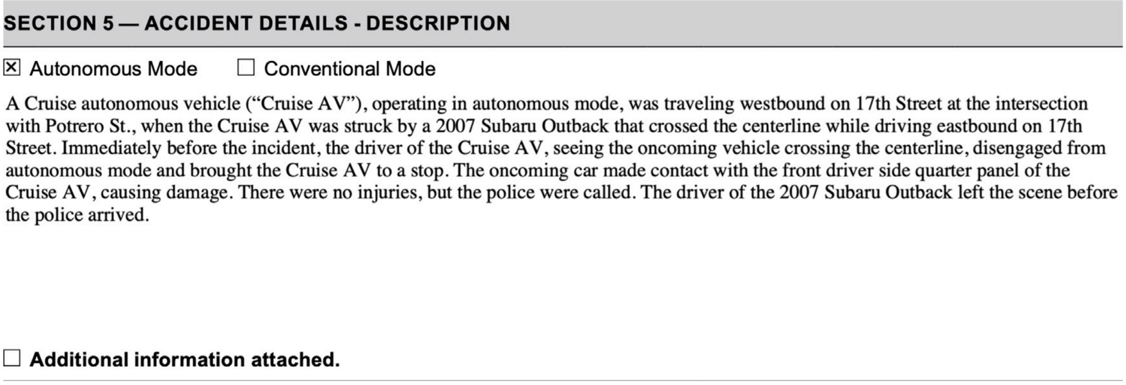
5.1.1. Model development

After data preparation, the generated vocabulary vector, the narrative matrix, and the metadata matrix are utilized as the inputs for developing the STM using the *R* package *stm* (Roberts et al., 2019). Multiple STM models have been developed using different numbers of topics, including 5, 10, 15, 20, 25, 30, and 35. To decide on an appropriate number of topics, metrics including residuals, exclusivity, and semantic coherence are utilized. An STM model that has a relatively small value of residual and large values of exclusivity and semantic coherence is preferred. Fig. 8 presents the metrics of STM models that have different numbers of topics. The model that has a number of topics of 15 is selected, as it reaches the lowest value of residual while having relatively high values of exclusivity and semantic coherence.

5.1.2. Topic interpretation

Interpreting topics in an STM model helps understand the latent information about AV crashes in crash narratives. Specifically, after determining the number of topics, Table 2 shows the top three words in probability, FREX, and Lift of individual topics. According to the words they contain, topics are divided into behavior-related, party-related, location-related, and general topics. Specifically, behavior-related topics contain words that depict the behavior of road users involved in crashes, including exchange, reverse, turn, yield, and stop. Behavior-related topics are the most common topics among all topics, which is consistent with reality as crash narratives usually contain descriptions related to the behavior of road users before, during, and after crashes. Party-related topics contain words that are related to parties involved in crashes, including scooters, bicyclists, and trucks. Location-related topics contain words that are related to locations where AV crashes occurred, such as intersections, expressways, etc. Lastly, general topics contain general words about AV crashes, such as the side, engage, minor, rear, etc. After converting crash narratives into topics, each crash narrative can be represented by the 15 topics with different proportions. In general, behavior-related crashes would have higher proportions of behavior-related topics. Similarly, location-related crashes would have higher proportions of location-related topics. This process converts the unstructured crash narratives into structured data, a N by 15 matrix, where N represents the number of crash narratives.

Another way of interpreting topics and their relationships with AV crashes is by estimating the proportions of topics in crash narratives. The mean topic proportion is used to show the overall topic distribution, which is estimated by averaging the topic proportion among all narratives as shown in Fig. 9. In addition, the top three words in terms of probability are annotated to help understand the meaning of individual topics. Topic 2 has the highest proportion among all topics, representing rear-end crashes, indicated by words including rear, bumper, and rear-end. This is consistent with the reality as rear-end crashes account for 51.6% of the total crashes. Topic 6 has the second-highest proportion and represents crashes that occurred on streets while involving cyclists. The word manual suggests that AVs may be in manual modes during the time of crashes or the drivers may disengage the autonomous modes into manual modes. Topics 12 and 9 have similar proportions. Topic 12 is a behavior-related topic that represents crashes involving lane-changing behavior. Similarly, Topic 9 is



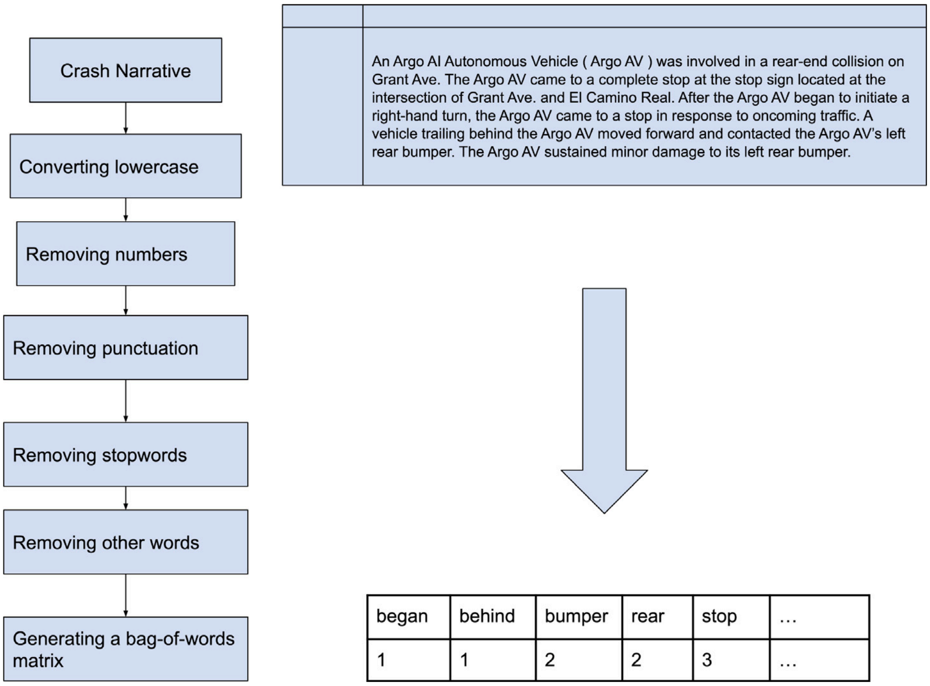
SECTION 5 — ACCIDENT DETAILS - DESCRIPTION

☒ Autonomous Mode ☐ Conventional Mode

A Cruise autonomous vehicle ("Cruise AV"), operating in autonomous mode, was traveling westbound on 17th Street at the intersection with Potrero St., when the Cruise AV was struck by a 2007 Subaru Outback that crossed the centerline while driving eastbound on 17th Street. Immediately before the incident, the driver of the Cruise AV, seeing the oncoming vehicle crossing the centerline, disengaged from autonomous mode and brought the Cruise AV to a stop. The oncoming car made contact with the front driver side quarter panel of the Cruise AV, causing damage. There were no injuries, but the police were called. The driver of the 2007 Subaru Outback left the scene before the police arrived.

☐ **Additional information attached.**

Fig. 4. An example of the crash narrative.



An Argo AI Autonomous Vehicle (Argo AV) was involved in a rear-end collision on Grant Ave. The Argo AV came to a complete stop at the stop sign located at the intersection of Grant Ave. and El Camino Real. After the Argo AV began to initiate a right-hand turn, the Argo AV came to a stop in response to oncoming traffic. A vehicle trailing behind the Argo AV moved forward and contacted the Argo AV's left rear bumper. The Argo AV sustained minor damage to its left rear bumper.

began	behind	bumper	rear	stop	…
1	1	2	2	3	…

Fig. 5. The process of preparing crash narratives.

Table 2
Topic-word distribution table.

Type	Topic	Probability	FREX	Lift
Behavior	8	left, exchange, pass	left, pass, exchange	white, pass, left
	9	park, reverse, mirror	park, reverse, mirror	lot, mirror, park
	11	turn, modern, van	turn, damag, modern	unprotect, begin, damag
	12	lane, short, chang	lane, bus, chang	encroach, rightmost, bus
	14	stop, behind, red-light	red-light, behind, forward	forward, acceler, red-light
	15	rear, yield, road	fascia, prepar, yield	fascia, lamp, prepar
Party	6	street, manual, cyclist	cyclist, street, supervise	cyclist, supervise, fell
	7	oncom, bicyclist, cross	bicyclist, scooterist, electr	electr, bicyclist, roadway
	13	right, truck, corner	corner, pedestrian, truck	corner, pedestrian, motorcyclist
Location	2	bumper, rear, rear-end	expressway, rear-end, bumper	cross, expressway, hatch
	10	intersect, front, enter	intersect, tire, enter	tire, intersect, green-light
General	1	side, door, wheel	door, well, wheel	well, door, occup
	3	proceed, struck, light	light, way, struck	autonomi, way, light
	4	engage, level, street	engag, level, suv	present, engag, level
	5	minor, rear, stop	minor, low, sign	low, start, minor

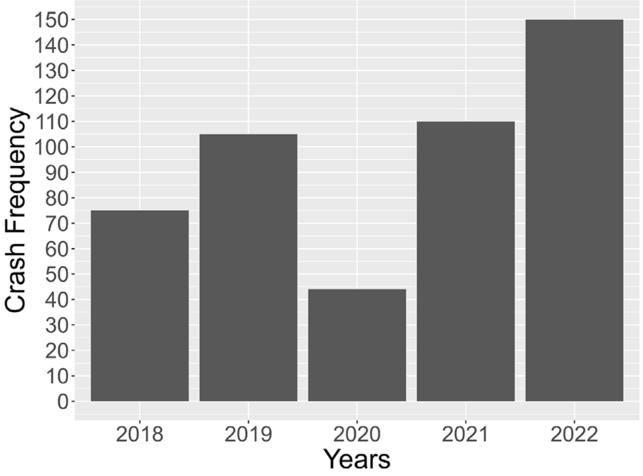


Fig. 6. The number of AV crashes from 2018 to 2022.

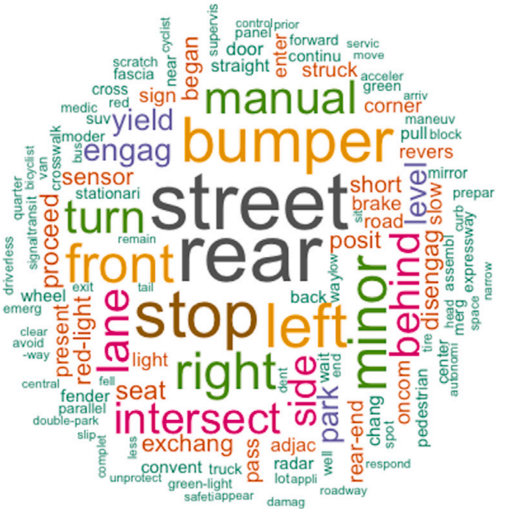


Fig. 7. The word cloud of the extracted crash narratives, the size of a word is positively related to its frequency in crash narratives.

a behavior-related topic representing crashes when vehicles are parked or in reverse. Overall, results about topic proportions suggest that rear-ended crashes are most common among AV crashes. Moreover, crashes that involve vulnerable road users require additional attention as they may result in more injuries, as suggested by the proportion of topic 6. In addition, AVs' ability to perform or respond to lane-change behavior needs to be further improved as topic 12 has a relatively high proportion among all topics.

In addition, to investigate the topic prevalence over time, the relationships between topic proportions and the year of crashes are estimated and presented in Fig. 10. Specifically, topic 2 becomes less prevalent over time, indicating that time negatively influences the proportion of topic 2. Topic 2, representing rear-end crashes, shows a gradual decrease in prevalence despite being the most common type of crash. For instance, rear-end crashes accounted for 61% of all crashes in 2018, declining to 50% by 2022. Meanwhile, topic 3 becomes slightly more prevalent over time, corresponding to the reality that the portion of crashes causing minor damage has increased since 2018. For example, the portion of crashes causing minor damage was 73% in 2018, which increased to 75% in 2022. In addition, time has a significantly positive impact on the proportion of topic 4, representing crashes associated with the disengagement behavior of AVs. Differently, the impact of time on topic 7 is relatively stable, suggesting that AVs' ability to interact with VRUs remains constant over the year but still requires improvements. Similarly, topic 8 has a relatively stable temporal prevalence, suggesting that lane-change behavior contributes consistently to AV crashes. This underscores the need for ongoing improvements in AV capabilities to execute and respond to lane-change maneuvers. In summary, results from interpreting topic prevalence over time suggest that although rear-end crashes still account for a large portion of crashes, their portions are decreasing over time. Differently, more and more AV crashes are causing minor damage over time. Additionally, AVs' capabilities in interacting with VRUs, performing, and reacting to lane-change maneuvers remain relatively constant over time, emphasizing the necessity for future enhancements.

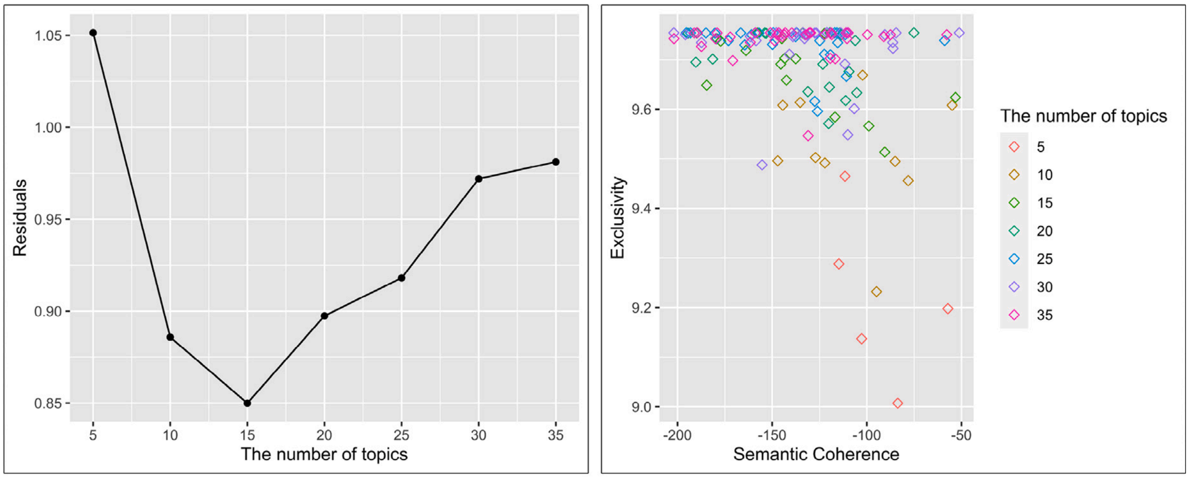
Table 3
Results of crash severity classification.

Data	Crash severity	Precision	Recall	F1-score
Training	None	0.994	0.999	0.997
	Minor	0.999	0.994	0.997
	Moderate and major	1.000	1.000	1.000
Test	None	0.945	0.933	0.939
	Minor	0.769	0.825	0.792
	Moderate and major	0.881	0.800	0.831

5.2. Crash severity modeling

The unstructured crash narratives are transformed into structured topics using the STM developed in the previous sections. The incorporation of metadata (e.g., crash severity) would improve the correlations between the topics and the crash severity. However, these relationships are still not clear as STM is an unsupervised model. Using the proportions of the topics as inputs, an XGBoost model is developed based on the methods presented in the previous section. Specifically, the narrative-topic proportion is prepared as a N by 15 matrix, where N represents the number of narratives. This matrix is used as the input of an XGBoost model that estimates the severity of crashes. The model's results on the training and test data are averaged over the five splits as shown in Table 3. The model has achieved promising accuracy on both training and test data with an average precision of 0.931, recall of 0.925, and F1-score of 0.926. The model has achieved superior performance compared to similar studies such as Zhu and Meng (2022), which suggests that the extracted topics are closely related to crash severity and it is necessary to consider the impact of metadata while developing a topic model.

Furthermore, SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) is used to interpret the relationships between topics and crash severity learned by the XGBoost. SHAP is a game theory-based method that assigns each feature an importance value to represent its contribution to the model's output (Lundberg and Lee, 2017). The SHAP values are additive, which means that the contribution of each feature to the final prediction can be computed independently (Li et al., 2024). After estimating SHAP values using the developed model, Fig. 11 presents the impact of topics on the mean absolute SHAP values. The larger the absolute value of the SHAP value a topic has, the more important it is for the model's outputs. Therefore, Fig. 11 suggests that topics 1, 3, 5, 6, 7, 8, 11, and 13 are the most important topics for identifying crash severity, while other topics have relatively less impact on the model's outputs. The reason is that the important topics represent specific information related to crash severity, while other topics represent relatively general information. For example, topic 2 has the highest value of mean proportion in Fig. 9, but its impact on the model is not as significant as other topics. The reason is that Topic 2 represents rear-end crashes, which are the most common crashes. Therefore, a general topic like it may not be very useful in terms of identifying crash severity. Topic 5, on the other hand, has a much smaller value of mean proportion than topic 2, but it is the most



(a) Residuals of different STM models

(b) Exclusivity and semantic coherence of different STM models

Fig. 8. Metrics of different STM models, (a) residuals, (b) exclusivity and semantic coherence.

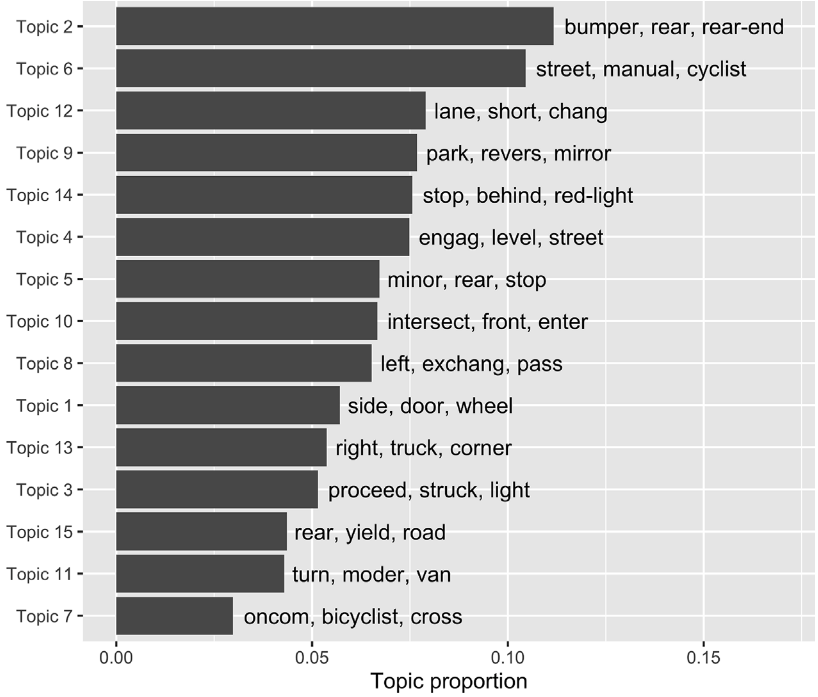


Fig. 9. The mean topic proportion among crash narratives, the top three words in terms of probability are annotated per topic.

important topic in terms of identifying crash severity. The reason is that topic 5 represents crash narratives that are related to minor injuries, therefore it is expected to have a significant impact on identifying crash severity. To further investigate the relationships between topics and crash severity, the impact of each topic on SHAP values of individual crash severity was estimated as shown in Fig. 12. Specifically, topic 5 has a strong positive impact on the probability of crashes causing minor injuries as suggested by Fig. 12(b). Furthermore, topic 7 has the smallest mean proportion among all crash narratives, but it is the most important topic for identifying crashes causing moderate or major injuries. Topic 7 represents crashes involving Vulnerable Road Users (VRUs) (e.g., bicyclists, scooters, etc.) and these crashes are more likely to cause serious injuries compared to crashes that only involve vehicles. Similarly, topic 6 represents crashes involving bicyclists and is therefore positively related to the probability of crashes causing moderate or major injuries. Finally, the characteristics of the crucial topics for crash severity classification are summarized as follows: topics 3, 5, and 11 provide specific information related to crash severity, such as light, minor, and moderate; topics 6, 7, and 13 provide specific information related to road users involved in crashes, including bicyclists, scooters, etc; topics 1 and 8 provide information about areas damaged on the vehicle and road user behavior, respectively. In summary, the topics extracted from STM have been proven to be closely related to crash severity, which further validates the effectiveness of STM in identifying latent information from AV crash narratives and the necessity of incorporating metadata while modeling crash narratives.

6. Conclusions

AVs are expected to bring huge potential benefits to the safety and efficiency of transportation systems, and they are considered to play an important role in reaching the Vision Zero goal. However, the development of AVs also raises concerns as AVs are still involved

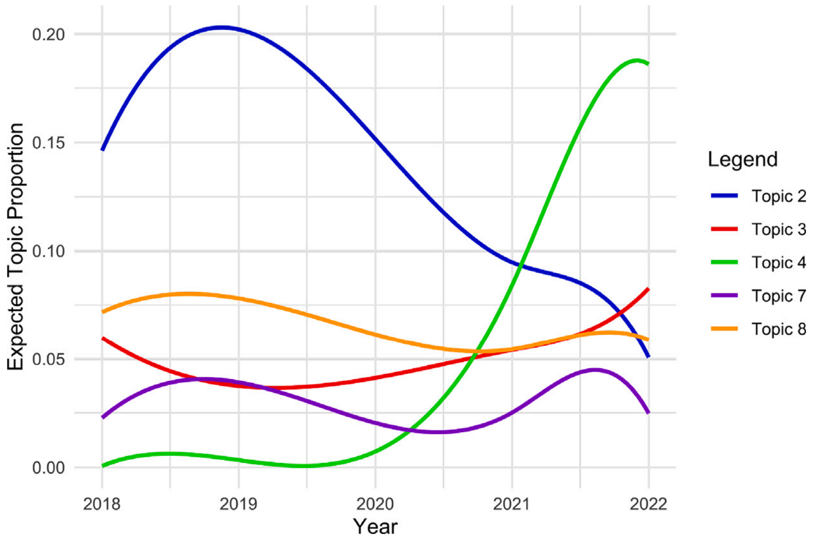


Fig. 10. Topic prevalence over time.

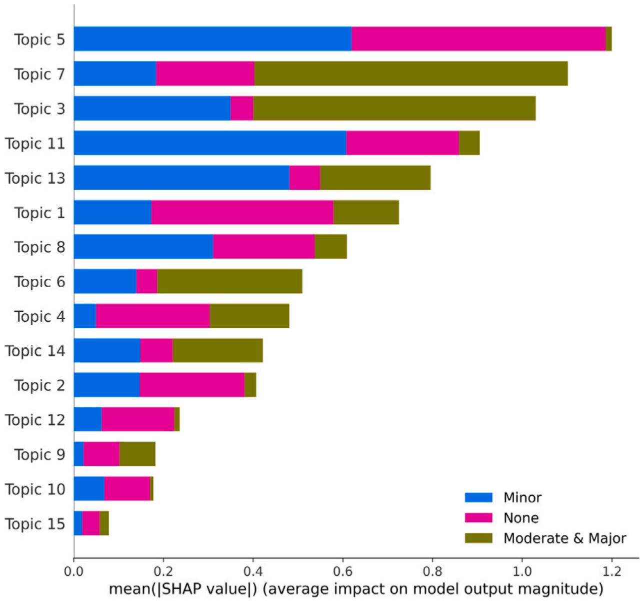
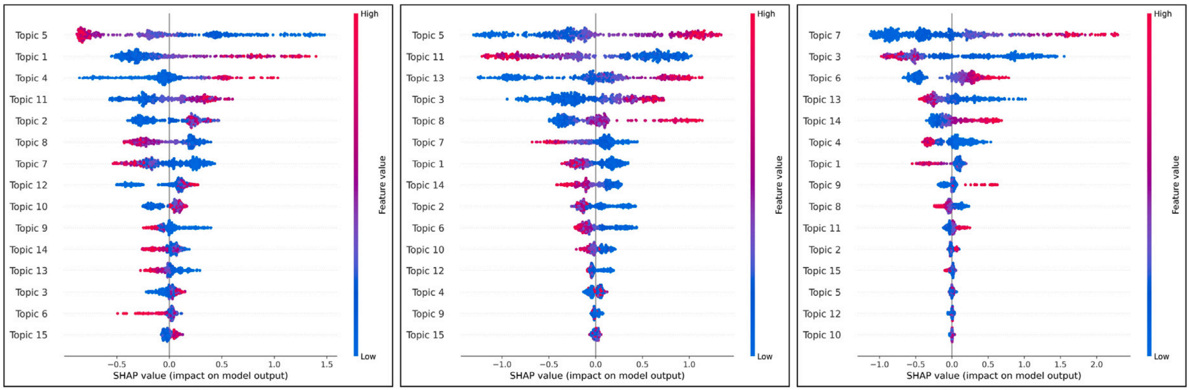


Fig. 11. The impact of topics on the mean absolute values of SHAP values.

in various crashes. Analyzing AV crashes could facilitate the understanding of their characteristics and contributing factors, thus providing suggestions for making safer AVs. Most existing studies have focused on the general descriptive statistics of AV crashes rather than analyzing the detailed crash narratives that may contain latent information about AVs. In addition, the contributing factors to AV crash severity require further investigation. Using AV crash data from 2018 to 2022 obtained through California DMV, this study converts the unstructured AV crash narratives into structured topics using a topic model (i.e., STM) by incorporating the severity and year of crashes. Furthermore, to understand the relationships between the extracted topics and crash severity, an XGBoost model is developed to identify crash severity using the proportions of topics. The model has achieved an average precision of 0.931, recall of 0.925, and F1-score of 0.926, suggesting that the extracted topics are closely related to crash severity. The conclusions of this study are summarized as follows:

- The topics among AV crashes could be divided into four types, including behavior-related, party-related, location-related, and general topics. Behavior-related topics account for most of the topics, indicating the close relationship between AV crashes and driver behaviors. Left-turning and lane-changing are especially prevalent among behavior-related topics. Moreover, VRUs, including pedestrians, cyclists, and scooters are prevalent among party-related topics, indicating that AV manufacturers should pay additional attention in terms of safety regarding VRUs.
- Results from topic-word distributions suggest that several topics have relatively large proportions compared to other topics. One of them is related to rear-end crashes, which account for more than 50% of AV crashes. One behavior-based topic that describes lane-change behavior also has a relatively high value of proportion, indicating the need to improve the AVs' ability to perform lane-change behavior or react to other road users' lane-change behavior.
- The relationships between the latent topics and crash severity are further analyzed by an XGBoost model, achieving high accuracy in terms of precision, recall, and F1-score. The XGBoost model is further interpreted using SHAP to comprehend these relationships. Results indicate that topics containing information about crash severity and VRUs have substantial impacts on the model's output compared to other topics. The proportion of a topic is not necessarily related to its importance in terms of identifying crash severity. For example, topics that contain general information about crashes may have higher proportions among crash narratives but are not essential for identifying crash severity.

In summary, this study has successfully extracted latent topics from AV crash narratives using STM while incorporating crash metadata including the year and severity of crashes. The extracted topics are closely related to the severity of crashes, suggested by both the results from XGBoost and SHAP. Results from this study indicate AV manufacturers should pay additional attention to AV's ability to interact with VRUs, perform lane-change behavior, and react to lane-change behavior of other road users. Furthermore, general information about crashes may not be enough to identify crash severity, specific information should be provided in AV crash narratives. In the future, additional research should be conducted using additional AV crash data as the current data only contains AV crashes in one state. Moreover, the information captured by AVs such as the behavior of surrounding road users, can be included as complementary data. The possibility of using the extracted topics to further classify AV crashes can also be explored.



(a) None **(b) Minor** **(c) Moderate & Major**

Fig. 12. The impact of topics on individual crash severity, (a) None injury, (b) Minor injury, (c) Moderate & major injury.

CRediT authorship contribution statement

Pei Li: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Sikai Chen:** Writing – original draft, Writing – review & editing, Conceptualization, Methodology, Supervision. **Lishengsa Yue:** Writing – original draft, Writing – review & editing. **Yuan Xu:** Validation, Visualization, Writing – review & editing. **David A. Noyce:** Conceptualization, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors would like to thank the California DMV for providing the data. All results and opinions are those of the authors only and do not reflect the opinion or position of the California DMV.

References

Alambeigi, H., McDonald, A.D., Tankasala, S.R., 2020. Crash themes in automated vehicles: A topic modeling analysis of the California department of motor vehicles automated vehicle crash database. arXiv preprint arXiv:2001.11087.

Bischof, J., Airoldi, E.M., 2012. Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning. ICML-12, pp. 201–208.

Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of science. Ann. Appl. Statist. 1 (1), 17–35.

Boggs, A.M., Wali, B., Khattak, A.J., 2020. Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical Bayesian heterogeneity-based approach. Accid. Anal. Prev. 135, 105354.

Brown, D.E., 2015. Text mining the contributors to rail accidents. IEEE Trans. Intell. Transp. Syst. 17 (2), 346–355.

California DMV, 2022. Autonomous vehicle collision reports. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/.

Chang, J., 2011. lda: Collapsed gibbs sampling methods for topic models.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Eisenstein, J., Ahmed, A., Xing, E.P., 2011. Sparse additive generative models of text. In: Proceedings of the 28th International Conference on Machine Learning. ICML-11, pp. 1041–1048.

Etherington, D., 2019. Over 1,400 Self-Driving Vehicles are Now in Testing by 80+ Companies Across the Us. Tech Crunch.

Favarò, F.M., Nader, N., Eurich, S.O., Tripp, M., Varadaraju, N., 2017. Examining accident reports involving autonomous vehicles in California. PLoS One 12 (9), e0184952.

Ghazizadeh, M., McDonald, A.D., Lee, J.D., 2014. Text mining to decipher free-response consumer complaints: Insights from the NHTSA vehicle owner's complaint database. Hum. Factors 56 (6), 1189–1203.

Grimmer, J., 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Polit. Anal. 18 (1), 1–35.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. C 44, 363–381.

Hu, J., Bhowmick, P., Arvin, F., Lanzon, A., Lennox, B., 2020. Cooperative control of heterogeneous connected vehicle platoons: An adaptive leader-following approach. IEEE Robot. Autom. Lett. 5 (2), 977–984.

International, S., 2021. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE Int. 4970 (724), 1–5.

Kwayu, K.M., Kwigizile, V., Lee, K., Oh, J.-S., 2021. Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. Accid. Anal. Prev. 150, 105899.

Leilabadi, S.H., Schmidt, S., 2019. In-depth analysis of autonomous vehicle collisions in California. In: 2019 IEEE Intelligent Transportation Systems Conference. ITSC, IEEE, pp. 889–893.

Li, P., Abdel-Aty, M., 2022. A hybrid machine learning model for predicting real-time secondary crash likelihood. Accid. Anal. Prev. 165, 106504.

Li, P., Wu, K., Cheng, Y., Parker, S.T., Noyce, D.A., 2024. How does c-v2x perform in urban environments? results from real-world experiments on urban arterials. IEEE Trans. Intell. Vehi. 9 (1), 2520–2530.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Advan. Neur. Info. Proc. Syst. 30.

Lv, C., Cao, D., Zhao, Y., Auger, D.J., Sullman, M., Wang, H., Dutka, L.M., Skrypchuk, L., Mouzakitis, A., 2017. Analysis of autopilot disengagements occurring during autonomous vehicle testing. IEEE/CAA J. Autom. Sin. 5 (1), 58–68.

Mimno, D.M., McCallum, A., 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In: UAI, vol. 24, Citeseer, pp. 411–418.

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 262–272.

NHTSA, 2020. Automated vehicles for safety. https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety.

NSTC, USDOT, 2020. Ensuring american leadership in automated vehicle technologies: automated vehicles 4.0. NSTC, USDOT: Washington, DC, USA.

Ooms, J., 2022. pdftools: Text extraction, rendering and converting of PDF documents. https://docs.ropensci.org/pdftools/ (website).

Roberts, M.E., Stewart, B.M., Airoldi, E.M., 2016. A model of text for experimentation in the social sciences. J. Amer. Statist. Assoc. 111 (515), 988–1003.

Roberts, M.E., Stewart, B.M., Tingley, D., 2019. Stm: An r package for structural topic models. J. Stat. Softw. 91, 1–40.

Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M., et al., 2013. The structural topic model and applied social science. In: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation. Vol. 4, Harrahs and Harveys, Lake Tahoe, pp. 1–20.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural topic models for open-ended survey responses. Am. J. Polit. Sci. 58 (4), 1064–1082.

Roque, C., Cardoso, J.L., Connell, T., Schermers, G., Weber, R., 2019. Topic analysis of road safety inspections using latent Dirichlet allocation: A case study of roadside safety in Irish main roads. Accid. Anal. Prev. 131, 336–349.

Shi, L., Qian, C., Guo, F., 2022. Real-time driving risk assessment using deep learning with XGBoost. Accid. Anal. Prev. 178, 106836.

Taddy, M., 2012. On estimation and selection for topic models. In: Artificial Intelligence and Statistics. PMLR, pp. 1184–1193.

Taddy, M., 2013. Multinomial inverse regression for text analysis. J. Amer. Statist. Assoc. 108 (503), 755–770.

Teoh, E.R., Kidd, D.G., 2017. Rage against the machine? Google's self-driving cars versus human drivers. J. Saf. Res. 63, 57–60.

Wang, S., Li, Z., 2019. Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data. Accid. Anal. Prev. 129, 44–54.

Zhu, S., Meng, Q., 2022. What can we learn from autonomous vehicle collision data on crash severity? A cost-sensitive CART approach. Accid. Anal. Prev. 174, 106769.