Journal of Safety Research 93 (2025) 335–341

Contents lists available at ScienceDirect

Journal of Safety Research

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

Improving the usability of large emergency 911 data reporting systems: A machine learning case study using emergency incident descriptions

N. Katherine Yoon[a], Tyler D. Quinn[b], Alexa Furek[c], Nora Y. Payne[d], Emily J. Haas[d,*]

[a] Centers for Disease Control and Prevention (CDC)/National Institute for Occupational Safety and Health (NIOSH). Pittsburgh, PA, United States
[b] West Virginia University. Morgantown, WV, United States
[c] FORMER CDC/NIOSH. Austin, TX, United States
[d] CDC/NIOSH, Pittsburgh, PA, United States

ARTICLE INFO

Keywords:
Data standardization
Emergency incident categorization
Automated classification
Natural language processing
Machine learning

ABSTRACT

Introduction: Emergency 9-1-1 incident data are recorded voluntarily within fire-department-specific computer-aided dispatch systems. The National Fire Incident Reporting System serves as a repository for these data, but inconsistency and variability in reporting practices across departments often lead to challenges in data quality and utility. This study aims to enhance emergency incident categorization and explore the feasibility of an automated system using free-text incident data from the National Fire Operations Reporting System (NFORS). *Method:* Researchers extracted and standardized 3,564 unique 9–1-1 incident descriptions from six fire departments using NFORS data, including narrative fields from emergency reports. The data were preprocessed using natural language processing (NLP) techniques, such as tokenization, stop word removal, and feature extraction (e.g., TF-IDF and n-grams). These features were used to train and evaluate Machine Learning (ML) models, including Naïve Bayes, Random Forest, and Support Vector Machine, to classify incidents into nine categories. The NLP techniques prepared the text data for the ML models, which performed the classification and assessed the automated system's performance. *Results:* The study demonstrated significant improvements in incident categorization accuracy using the NLP and ML approach. Unigram models achieved 93% accuracy when applied to 3,564 unique incident descriptions. This performance was evaluated by comparing the automated classifications to manually assigned categories, which served as the reference. Mis-categorizations primarily occurred with "Emergency Medical Services (EMS)." *Conclusions:* Standardized and consistent incident categorization is vital for informed decision-making, efficient resource allocation, and effective emergency response. Our findings suggest that adopting a robust categorization system, such as the nine-category model using NLP and ML, can improve categorization accuracy and enhance data quality and utility for decision-making. *Practical Applications:* Public safety agencies can leverage these insights to modernize data systems, strengthen occupational surveillance, and create more resilient and sustainable public safety data systems.

1. Introduction

The details of emergency 9-1-1 incident calls are voluntarily recorded in fire-department-specific computer-aided dispatch (CAD) systems or record management systems (RMS). Such data entry systems often use the National Fire Incident Reporting System (NFIRS) to characterize information about each incident such as the time, date, type, response time, civilian and fire fighter casualties, property loss and other relevant details. The U.S. Fire Administration (USFA) established NFIRS as a voluntary reporting system in 1975. NFIRS was initially launched with six states in 1976 and is in its fifth edition (National Fire Data Center, 2017). A primary function of NFIRS is to serve as a repository for data related to fire and emergency response incidents across the United States (National Fire Data Center, 2015).

Over the years, NFIRS has evolved to encompass various aspects of fire incidents, including fire suppression, emergency medical services (EMS), hazardous materials responses, and fire-related deaths and injuries. Despite its comprehensive scope, NFIRS has faced major challenges, including its reliance on voluntary reporting, inconsistent use by dispatchers, as many as 567 data elements, and variations in how

* Corresponding author at: 626 Cochrans Mill Rd., Pittsburgh, PA 15236, United States.
E-mail addresses: NYoon@cdc.gov (N.K. Yoon), tyler.quinn1@hsc.wvu.edu (T.D. Quinn), Amf180@pitt.edu (A. Furek), ttt2@cdc.gov (N.Y. Payne), wcq3@cdc.gov (E.J. Haas).

https://doi.org/10.1016/j.jsr.2025.04.001
Received 4 December 2024; Received in revised form 3 March 2025; Accepted 1 April 2025
Available online 7 April 2025
0022-4375/Published by Elsevier Ltd.

different fire departments record incidents, sometimes leading to incomplete or unreliable data (National Fire Data Center, 2017). Considering the limitations of NFIRS, the USFA announced its commitment to data modernization in the development of a new, interoperable fire information and analytics platform labeled the National Emergency Response Information System (NERIS) (U.S. Fire Administration, 2023). Once developed, piloted, and refined, NERIS will eventually replace NFIRS to provide the firefighter community with empirical, reliable data analytics.

Meanwhile, the National Fire Operations Reporting System (NFORS), developed by the International Public Safety Data Institute (IPSDI), offers a modern approach to incident reporting (IPSDI, 2023). While NFORS follows a similar structure to NFIRS, it streamlines incident category by grouping them into three large categories: “EMS,” “Fire,” and “Other.” This categorization is based on the description of the incident and the type of responder (e.g., fire unit only or fire-based EMS unit). NFORS automates the extraction, standardization, and analysis of incident data directly from fire department CAD or RMS, making it efficient and less prone to the inconsistencies.

This study leverages NFORS data and employs broadly applied modern data science techniques to explore potential ways to improve the usability of NFIRS data with the intent to inform future NERIS data reporting elements and to provide insights into data standardization. This study included two primary objectives: (1) identify an updated, more accurate, and informative incident scheme to categorize data; and (2) examine the feasibility of an automated incident category system based on free-text analysis of incident data.

Methods.

As part of a project within the Centers for Disease Control and Prevention (CDC) Data Modernization Initiative (Centers for Disease Control and Prevention. Data Modernization Initiative, 2023), the National Institute for Occupational Safety and Health (NIOSH) collaborated with IPSDI, a non-profit data science organization that procures and analyzes public safety surveillance data (IPSDI, 2023). IPSDI provided NIOSH with 9-1-1 emergency incident descriptions after they were normalized within IPSDI’s proprietary NFORS, which is different from NFIRS but provides an integrated approach to data extraction and standardization for fire departments. The decision to use NFORS in this study was driven by its capability to provide normalized 9-1-1 emergency incident descriptions, which are essential for analyzing and improving incident categorization. For fire departments that engage with IPSDI, NFORS is integrated directly into the department’s CAD or RMS for automatic extraction and standardization of operational data. This activity was reviewed by CDC, deemed research not involving human subjects, and was conducted consistent with applicable federal law and CDC policy.[1]

The data sample received from IPSDI included 3,511,066 incidents that occurred between 2016 and 2021 from six fire departments. When the incident data were retrieved from department systems, IPSDI standardized the incidents, assigning each to one of three categories: *EMS, Fire*, or *Other*. The researchers initially generated word clouds of the incident narrative descriptions by each of the three incident categories (Fig. 1) to serve as a starting point for further analysis of the data (Viegas et al., 2007; Sinclair & Cardew-Hall, 2008; Cidell, 2010; DePaolo & Wilkinson, 2014). These word clouds provided a visual representation of the most frequently occurring terms within each category. Specifically, the bigger and bolder each word is, the more often it appeared in the dataset, helping to highlight key themes and patterns in the incident descriptions, which informed the categorization process.

Fig. 1 shows inconsistencies in the categorization for some incidents. For example, *person, ill, injured, pain*, and *unconscious* were all keywords accurately assigned to an *EMS* incident in the word cloud. However, the *Other* word cloud shows *stroke* and *unconscious,* which should be categorized as *EMS* incident keywords. Also, because there are potentially many more categories of incidents than the three described, some incidents were forced into nonrelevant categories. For example, *motor vehicle accidents* were categorized as *Fire* regardless of the presence of information that could indicate whether those accidents involved a fire. Moreover, some incidents were clearly coded incorrectly.

Based on these initial mis-categorizations, the researchers sought to accomplish two objectives: (1) identify an updated, more accurate, and more informative incident scheme to categorize data; and (2) examine the feasibility of an automated incident category system based on the free-text incident data.

1.1. *Review and selection of the data sample*

First, researchers systematically reviewed the full dataset of approximately 3.5 million incident descriptions to identify and remove duplicate incident descriptions. The primary reason for eliminating duplicates was to prevent overfitting in ML models. This process resulted in the extraction of 3,564 unique incident descriptions. Within this set of unique incidents, some descriptions were still qualitatively similar. For example, variations like “fall-unconscious” and “fall-unconscious-public” or “fall-unknown status-ground” and “fall-unknown status-public” were included as unique entries, even though they convey essentially the same information about incidents. To address this issue, researchers conducted a thorough examination of these qualitatively similar incident descriptions. This involved comparing their wording, context, and specificity of each entry to determine whether they represented distinct incidents or redundant variations. Researchers also assessed the potential impact of merging similar incident descriptions on the dataset’s clarity, specificity, potential misclassification, and implications for subsequent analyses. After careful examination, the researchers chose not to merge these qualitatively similar descriptions, for reasons described below.

First, maintaining the raw data’s authenticity was crucial for effectively training the ML models. Minor variations in incident descriptions, even those seemingly insignificant, may carry contextual nuances that are valuable for models to learn and generalize. Further, the preservation of these distinctions allowed the models to adapt to the subtleties of real-world language, making them more robust and capable of handling a wider range of scenarios. Finally, this approach ensured that the analysis remained true to the unaltered raw data from emergency response calls, providing a comprehensive and unmanipulated perspective on the incidents encountered.

Because the original categories (*EMS, Fire*, and *Other*) in NFORS were quite broad and designed for operational efficiency, the researchers recognized that these categories alone might not provide sufficient detail to accurately understand service utilization rates, national incident burdens, and other critical aspects. NFORS uses a simplified set of categories to facilitate data extraction and standardization. To gain a more nuanced understanding, the team consulted the NFIRS incident categories and broadened IPSDI’s three category coding system to include more options, using insights from the word clouds shown in Fig. 1 (e.g., motor vehicle accidents, hazmat-related incidents, and overdoses were all visible). The researchers used this broader set of incidents during manual recoding of the unique 3,564 incident descriptions to address the first objective of the study.

Objective 1: Identifying an Informative Incident Category Scheme.

First, a more detailed review of the data and separate coding was necessary to ensure that incident categories appropriately and accurately matched the incident descriptions (i.e., the narrative information provided in the dispatch report). To determine the accuracy of incident categories and opportunities for improvement, researchers manually coded the data, using an Excel spreadsheet, for the 3,564 unique incident descriptions.

Researchers split the data file into two parts, each containing 1,782 incident descriptions. Then, four researchers were divided into two

[1] See e.g., 45C.F.R. part 46; 21C.F.R. part 56; 42 U.S.C. §241(d), 5 U.S.C. §552a, 44 U.S.C. §3501 et seq.

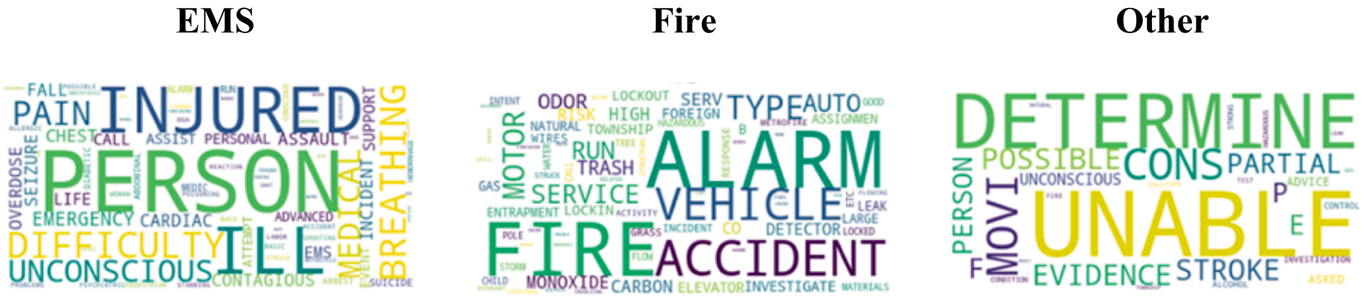

Fig. 1. Word clouds for original narrative descriptions of 3,511,066 incidents.

teams of two researchers, with each team independently reviewing and coding one of the two parts. Within each team, the two researchers coded their part independently. The researchers assigned a category to each incident using the following options: *EMS*, *Fire*, *Hazmat/Explosive*, *Motor Vehicle Incident*, *Overdose*, *Psychological*, *Public Service Assistance*, *Contagious Emergency*, and *Other/Unclear*. The options were derived from some overarching categories in NFIRS. After they finished coding independently, the four researchers met to further discuss and adjudicate any discrepancies. The purpose of this collective effort was to ensure a robust and consistent categorization of the data.

Through open dialogue and mutual agreement, the researchers deliberated on each case, considered different perspectives, and arrived at a consensus regarding the most appropriate category for each instance. This collaborative approach aimed to enhance the reliability and validity of the coding process, fostering a shared understanding among the research team. Subsequently, ML models were run to evaluate the performance of the initial categorization. Instances of misclassification were analyzed to determine whether these errors were due to flaws in the categorization process or model limitations. It is important to note that all nine categories were kept and no additional data categories beyond those manually coded were considered during this evaluation. The manual coding was used as the gold standard for comparison. This analysis led to the re-coding of approximately 50 cases to correct inaccuracies identified in the initial categorization.

Table 1 provides examples of incident descriptions with their original and recoded categories. For instance, automatic alarms initially categorized as *Fire* were reclassified as *Other/Unclear* when insufficient contextual information was available. Similarly, incidents of suicide attempts and emotionally disturbed persons were recoded from *EMS* and *Other* to *Psychological*, respectively.

To assess the improvements achieved by transitioning from the original three category system to the new nine category scheme, researchers performed a comparative analysis. We identified 4,117 uniquely mapped incident description-category pairs (not 3,564) from the original three category dataset. This analysis highlighted the inconsistencies and variations in the original categorization, as the same incident descriptions were assigned different categories.

Table 1
Examples of incident descriptions with original and updated category assignments.

Incident description	Original category assigned by NFORS	Updated category assigned by researchers
x77c02i—fuel/fluid leak	EMS	Hazmat/Explosive
motor vehicle collision	EMS	Motor Vehicle Incident
poisoning/overdose (ingestion)	EMS	Overdose
psych emergency jumper—violent	EMS	Psychological
contagious emergency	EMS	Contagious Emergency
aslt—no danger proximal sexual	EMS	Public Service Assistance
automatic alarm	Fire	Other/Unclear
elevator service run	Fire	Other/Unclear
lockout lockin	Fire	Public Service Assistance
fire alarm residential	Other	Fire
water/ice/mud rescue	Other	Other/Unclear
unconscious person (e) (f) (p)	Other	EMS
emotionally disturbed person acting aggressively	Other	Psychological

Objective 2: Examining the Feasibility of an Automated Incident Category System.

After extracting the narrative incident descriptions and re-categorizing them using the updated nine categories, researchers employed a methodology combining natural language processing (NLP) techniques and ML algorithms (Naïve Bayes, Random Forest, and Support Vector Machine) to categorize all incidents using the nine categories. Prior to modeling, several preprocessing steps enhanced the quality and consistency of the narrative incident data. All analyses used Python version 3.9.12.

Data Preparation and Method.

Preprocessing steps helped refine the incident narratives and alleviate noise or irrelevant information, preparing the data for subsequent analysis and categorization using NLP and ML algorithms (Bird et al., 2009; Gupta & Lehal, 2009; Jurafsky & Martin, 2000; Manning & Schutze, 1999). Preprocessing included three steps:

1. Removed common stop words (such as “the,” “and,” or “is”) and punctuation that frequently appeared in the incident narratives but did not carry significant meaning or contribute to the category.
2. Applied case normalization to convert all text to lowercase to ensure consistency.
3. Employed lemmatization to reduce words to their base or root forms, consolidating variant forms of words.

Following the preprocessing steps, researchers utilized two fundamental NLP techniques to further prepare the data for modeling. The first technique was Term Frequency-Inverse Document Frequency (TF-IDF) (Baeza-Yates & Ribeiro-Neto, 1999; Manning & Schutze, 1999; Zhai & Lafferty, 2004). The second was n-gram representations with different n-gram lengths (unigrams, bigrams, and trigrams) (Suen, 1979; Manning & Schutze, 1999; Nadkarni et al., 2011).

As seen in Table 2, another notable characteristic of the dataset was its inherent class imbalance. Across the complete dataset, the researchers observed a disparity in the distribution of incident categories. In particular, among the 3,564 unique incident descriptions, the *EMS* category emerged as the majority class, representing approximately 58.2% of the data, followed by the *Motor Vehicle Incident* category at 9.5%, and the *Fire* category at 7.6%. The remaining six categories collectively constituted the remaining 24.7%, with individual categories ranging from 0.2% (*Contagious Emergency*) to 6.8% (*Other/Unclear*). Despite the class imbalance, the researchers opted not to artificially balance the dataset to maintain the real-world distribution of incident categories. Our results (see Results section and confusion matrix in Appendix Fig. A1) do not suggest that the level of class imbalance in the data had a substantially negative impact on classifier training and

Table 2
Comparison of the number and percentage of each category among unique incident descriptions (N = 3,564) and among all incident descriptions (N = 3,511,066).

Category	# of 3,564 unique incident descriptions	% of 3,564 unique incident descriptions in each category	# of incidents (all)	% of incidents in each category (all)
EMS	2,075	58.2%	2,334,419	66.5%
Fire	271	7.6%	313,754	8.9%
Hazmat/ Explosive	202	5.7%	62,432	1.8%
Motor Vehicle Incident	338	9.5%	279,621	8.0%
Overdose	181	5.1%	61,709	1.8%
Psychological	151	4.2%	50,091	1.4%
Public Service Assistance	95	2.7%	132,254	3.8%
Contagious Emergency	8	0.2%	11,424	0.3%
Other/Unclear	243	6.8%	265,362	7.6%
TOTAL	3,564	100%	3,511,066	100%

performance.

Also, a potential concern when performing analyses on data from which repeated incident descriptions have been removed is that the distribution of incident types may differ significantly from the full dataset. If the distribution of incident types in the deduplicated and sample datasets are vastly different, then the results on the deduplicated dataset may not accurately reflect the results of applying the model in a real-world scenario. Table 2 suggests that this is not a major concern, as the distribution of incident types in the deduplicated dataset is similar to that in the full dataset. Consequently, researchers felt that resampling techniques were not necessary to restore the class balance reflected in the full dataset to the deduplicated data.

Supervised Machine Learning Models.

Researchers employed Naïve Bayes, Random Forest, and Support Vector Machine (SVM) models in this study. Naïve Bayes is a simple model assuming independent features and is effective for text classification. Random Forest is an ensemble method that builds multiple decision trees and combines their results to improve accuracy. The focus of SVM is on finding the best boundary to separate different categories, even in complex datasets. Each algorithm has its own strengths in handling different types of data and tasks (Hearst et al., 1998; Breiman, 2001; Rish, 2001).

Multiple algorithms allowed researchers to assess and compare model category effectiveness and performance. To evaluate the performance and generalization capability of the models, researchers adopted a rigorous model training and evaluation methodology. After dividing the preprocessed incident data into training (75%) and testing (25%) sets, researchers trained the Naïve Bayes, Random Forest, and SVM models on the training set. To see if explicitly addressing class imbalances in the dataset would improve performance, a weighted Random Forest model using unigrams was also trained to compare it to its unweighted counterpart. Default parameter settings were used in most cases, with the following exceptions: Random Forest (n_estimators = 200); SVM (kernel= 'linear'). The trained models were then applied to the test set. Several performance metrics were calculated (see Table 3 for accuracies and F1 scores). Accuracy is the proportion of correct predictions made by the model. F1-score is the harmonic mean of precision and recall. It is especially useful when working with imbalanced datasets (Sokolova & Lapalme, 2009; Powers, 2020). Precision measures how many of the predicted positive instances are truly positive, while Recall (or Sensitivity) shows how many of the actual positive instances were correctly predicted as such. Performance metrics were rounded to two decimal places (equivalently, the nearest whole percentage).

Table 3
Test set results for the ML algorithms.

	Naïve Bayes		Random Forest		SVM	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Unigram	0.85	0.61	0.93	0.90	0.92	0.89
Uni + Bigram	0.80	0.53	0.92	0.90	0.92	0.88
Uni + Bi + Trigram	0.75	0.44	0.92	0.89	0.92	0.88

2. Results

Models using unigram features consistently demonstrated higher accuracy compared to models using unigram and bigram features or models using unigram, bigram, and trigram features (Table 3). Across all models, the Random Forest model achieved the highest accuracy of 93%, closely followed by SVM with 92% accuracy and Naïve Bayes with 85% accuracy. The F1 scores of these models reflected a similar pattern on test data (Random Forest: F1 = 0.90; SVM: F1 = 0.89; Naïve Bayes: F1 = 0.61). As expected, these results resembled 5-fold cross-validation results that the researchers had also calculated (see Table A1).

Examination of misclassified cases revealed that most of these instances involved incidents from other categories being inaccurately predicted as *EMS* (confusion matrix and classification report by category with Random Forest model using unigrams appear in the Appendix). Specifically, there were instances where cases categorized as *Other/Unclear* were misclassified as *EMS*. Examples such as "routine ambulance re" or "ee" were expected to be classified as *Other/Unclear* but were misclassified as *EMS*. This suggests that the model has a lower accuracy in predicting cases as *Other/Unclear* and tends to predict them as *EMS*, likely due to the dominance of *EMS* cases in the 3,564 dataset. Note that for *EMS*, the Random Forest model using unigrams had 97% precision, 96% recall, and a 97% F1 score. Across all categories, precision ranged from 68% (*Other/Unclear*) to 100% (*Psychological* and *Contagious Emergency*) while recall ranged from 73% (*Psychological*) to 100% (*Overdose*, *Contagious Emergency*). Misclassifications were predominantly observed in cases where incidents were predicted as *Other/Unclear* or *EMS* instead of other categories.

Tables 4 and 5 presents classification reports for the unweighted and weighted Random Forest models. The overall accuracy of both models was similar (93% compared to 94%). For most categories, the precision, recall, and F1 score was also similar between the unweighted and weighted models in most cases. For several categories, the weighting appeared to have a negative effect, for example, the precision and F1 score for *Public Service Assistance*.

3. Discussion

This study had two objectives: (1) identify an updated, more accurate, and more informative incident scheme to categorize data; and (2) examine the feasibility of an automated incident category system based

Table 4
Classification report by category for unweighted Random Forest model using unigrams.

	Precision	Recall	F1-score	Support
Contagious Emergency	1.00	1.00	1.00	1
EMS	0.97	0.96	0.97	534
Fire	0.83	0.89	0.86	62
Hazmat/Explosion/Gas/Bomb	0.94	0.92	0.93	53
Motor Vehicle Incident	0.97	0.88	0.92	88
Other/Unclear	0.68	0.84	0.75	69
Overdose	0.97	1.00	0.99	38
Public Service Assistance	0.89	0.73	0.80	11
Psychological	1.00	0.83	0.91	35
Accuracy			0.93	891
Macro average	0.92	0.89	0.90	891
Weighted average	0.94	0.93	0.93	891

Table 5
Classification report by category for weighted Random Forest model using unigrams.

Category	Precision	Recall	F1	Support
Contagious Emergency	1.00	1.00	1.00	1
EMS	0.98	0.97	0.97	534
Fire	0.84	0.87	0.86	62
Hazmat/Explosion/Gas/Bomb	0.94	0.96	0.95	53
Motor Vehicle Incident	0.99	0.90	0.94	88
Other/Unclear	0.70	0.86	0.77	69
Overdose	0.97	1.00	0.99	38
Public Service Assistance	0.80	0.73	0.76	11
Psychological	1.00	0.80	0.89	35
Accuracy			0.94	891
Macro average	0.91	0.90	0.90	891
Weighted average	0.94	0.94	0.94	891

on free-text analysis of incident data. To accomplish these objectives, this study utilized NLP methods, such as TF-IDF and n-grams, in combination with ML algorithms to categorize 9-1-1 incidents based on their narrative descriptions. The observed categorization errors highlighted the challenges faced in accurately assigning incidents to their respective categories, particularly when incident descriptions share similarities in the narratives or when an incident category dominates the sample. While these errors represent limitations of the current model, it is important to note that overall accuracy rates remained high. Specifically, the Random Forest model using unigrams achieved a 93% test accuracy, the best accuracy among all models included in this study. These results underscore the potential for improving emergency response standardization through data modernization efforts.

The data analysis and the implications of the results are unique in nature. Specifically, most analyses of emergency response calls focus on analyzing EMS data for specific trends in illnesses such as strokes, opioids, and SARS-CoV-2 (Friedman et al., 2021; Lasher et al., 2019; Garg et al., 2019; Pitt, 2022). However, there is still a lack of focus on unifying disparate datasets via data standardization, contributing to inconsistent and error-prone response efforts, as observed by Kim and colleagues (2021). Kim and colleagues (2021) went on to indicate that, to improve emergency response guidelines, "a significant amount of manual effort and domain expertise is needed for labeling and making sense of such data" (pg. 1). This study addressed this challenge by establishing a more refined categorization system and helping to ensure the feasibility of analytical endeavors.

Practical implications

As indicated earlier, NFIRS is a comprehensive database that enables the fire service community, policymakers, and researchers to gain insights into fire-related incidents, their causes, and the effectiveness of emergency response strategies. Although NFIRS has been a valuable tool for understanding fire trends, identifying risk factors, and guiding the development of fire prevention and response policies, it does not mandate or represent a set standard to report emergency incidents or code the narrative incident descriptions received. This lack of structure, in combination with a decentralized public health and safety reporting system, has been an ongoing challenge to sharing information across various data systems—resulting in the need for more modern EMS platforms (Henderson, 2009) that can improve interoperability (Valecha et al., 2013).

More recently than in the above studies, this challenge was apparent during the COVID-19 response when critical gaps were revealed in public health and safety data infrastructure that called for improved data surveillance and modeling (Hoff et al., 2023; Pitt, 2022). Primarily, the need for expanded and interoperable systems to inform real-time, accurate decision-making among national, state, and local entities became more critical. While it was not practical to scale such systems during the emergency response, there are current opportunities to improve data management and modernization including 9–1-1 incident reporting.

Researchers in the present study sought to improve the usability of NFIRS data with the intent to inform future NERIS data reporting elements. These elements would function as a national hub for data collection, standardization, and use in the emergency response realm. The application of ML to incident data categorization is not limited to the NFIRS database though. The methods employed in this study could be extended to other public health and emergency response datasets, including injury prevention fields. For example, ML has been used on large sets of worker compensation data at company and state levels to identify risk factors for work-related musculoskeletal disorders (Chan et al., 2022) while other studies have used a combination of large lagging indicator datasets to understand contributing factors and leading indicators for preventing more serious workplace injuries (Pramanik & Modak, 2024). The unique impact of this study and application of ML, however, is that validated ML models can hopefully be used in real-time in a way that befits the organization as emergency data is ingested into the surveillance system.

For example, although this study used ML to focus on categorization, further exploration of the newly categorized data could reveal trends by year, fire department, or time of day, potentially improving response times, resource allocation, and public safety strategies. These analyses, while beyond the scope of this study, offer valuable opportunities for future research to enhance emergency response decision-making to prevent injuries.

As other research has demonstrated, the current results emphasize the importance of refining and standardizing incident coding. This study identified nine categories of incidents to better understand the range of emergency incidents that occur and are dispatched to responders (see Fig. A2 for an updated word cloud). Specifically, further recognizing various subcategories of *EMS* incidents may enhance future dispatcher and fire department coding and improve the accuracy of categorizing incidents moving forward.

These results support the use of certain incident codes that occur more frequently in the data but were not adequately or specifically accounted for in department coding (e.g., psychological incidents). Although the differences may seem minor, specifically recognizing and coding for such incidents has important implications. First, these variations determine whom to deploy to a response, and second, they highlight the need for additional budgeting and resource support in a specific health or safety category. In the above example, specifically tracking psychological incidents can aid in identifying the need for specialized training for responders to effectively engage with individuals experiencing such events. Other research has also identified the value of standardizing incident descriptions and coding to better predict and allocate the need for personal protective equipment during supply shortages (Haas et al., 2021; Haas et al., 2023).

It is noteworthy to acknowledge that some incident descriptions such as "ee" occurred exclusively in Tennessee, while "routine ambulance re" was specific to Florida within our dataset. These occurrences may not necessarily be mistakes but rather reflect regional or department-specific conventions or terminology. This suggests that further training or adjustments to the model may be needed to account for such department-specific terms, potentially improving the accuracy of future predictions. Expanding the dataset to incorporate a wider variety of incident descriptions from different regions and departments could help the model better generalize and manage regional terminology variations. Also, continuous model evaluation and updates based on new data, along with incorporating expert feedback for validation, will ensure ongoing accuracy and relevance.

Limitations

The findings highlight the potential of NLP techniques coupled with ML algorithms to effectively categorize emergency incidents, thus contributing to the potential generation of standardized, comprehensive, and timely public safety surveillance data. However, several limitations must be acknowledged. First, manual coding and recategorization of the incident descriptions introduced the possibility of bias. To mitigate this, the researchers employed a multi-coder system, involving four coders to independently code and categorize the same data, followed by discrepancy adjudication among the four coders when necessary. Nevertheless, challenges arose when encountering unknown words, leading to coding uncertainties addressed through educated guesses and extensive discussions.

It is also important to reiterate that a substantial portion of the data were duplicates. Of those duplicates, although they were different quantitatively, they were not all qualitatively different as illustrated in examples throughout the paper. Furthermore, the dominance of *EMS* incidents in the dataset influenced the models to favor *EMS* incidents, potentially impacting the overall performance. To address these limitations in the data, researchers adopted evaluation metrics and methodologies tailored to address the nuances of class imbalance, underscoring a commitment to a comprehensive and informed analysis. This included utilizing precision, recall, and F1 score metrics at each category level (Tables 4 and 5), as well as analyzing confusion matrices (Fig. A1).

Along a similar vein, the data for this study were for emergency incidents that occurred during the COVID-19 pandemic at a time when many dispatch centers temporarily included a call type coded “contagious emergency” (Anderson, 2020). This ad-hoc call type was documented in many CAD systems across the country and the NFORS database extracted and included this call type as well. Consequently, including *Contagious Emergency* as a category in this specific study was warranted. While the inclusion of *Contagious Emergency* as a category was directly tied to the COVID-19 pandemic and may no longer identified as a call type in a more current version of the dataset, this example underscores the broader importance of having developed, validated models that can adapt to new, evolving public health emergencies or emerging contagious diseases. Such models should be capable of identifying new risk factors and enabling the continuous updating of prevention methods and resource allocation as new challenges arise. However, if the same analysis was completed on a more current version of the dataset, it is likely that *Contagious Emergency* would not emerge as its own category. This example shows the importance of having the capability to have developed, validated models that can be used on new, big datasets to identify contributing risk factors and update prevention methods and focused resources as needed.

Finally, the study revealed discrepancies in coding and language used across six fire departments, indicating that new data from different departments may introduce significant variations not captured in the currently presented models. This would require ongoing monitoring and management. In such cases, manual coding processes are recommended to ensure consistent coding with existing data until any significant differences are identified and addressed.

4. Conclusions

This study demonstrates that NLP and ML techniques are effective tools for categorizing emergency incident narratives. The results underscore the value of adopting a more comprehensive nine-category system for incident reporting, which enhances data accuracy and supports better resource allocation and decision-making.

The findings also highlight that while the current automated classification system shows promising results, ongoing improvements and expansions are needed. Future research should focus on incorporating a broader range of data sources and addressing regional variations to further refine and validate the classification system. This study provides a foundational step towards developing a more robust and standardized approach to emergency incident categorization.

5. Disclaimer

The findings and conclusions are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention. Mention of any company or product does not constitute endorsement by NIOSH, CDC.

CRediT authorship contribution statement

N. Katherine Yoon: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Tyler D. Quinn:** Formal analysis, Data curation. **Alexa Furek:** Data curation. **Nora Y. Payne:** Writing – review & editing. **Emily J. Haas:** Writing – review & editing, Project administration, Data curation.

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.

Acknowledgement

This work was supported by data from the National Fire Operations Reporting System (NFORS), which was provided by the International Public Safety Data Institute (IPSDI) and its contributing members for research, education, standards development, and emergency response purposes.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsr.2025.04.001.

References

Anderson, A. (2020). Case Study: Coronavirus (COVID-19) Dashboard in NFORS/StatEngine. Published in IPSDI’s Medium Webpage. March 19, 2020. https://medium.com/ipsdi/case-study-coronavirus-covid-19-dashboard-in-nfors-statengine-613928d0c267 (accessed January 2025).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463, No. 1999). New York: ACM press.

Bird, S., Klein, E., & Loper, E. (2009). *Categorizing and tagging words* (p. 179). In Natural Language Processing with Python: O’Reilly Media Inc.

Breiman, L. (2001). *Random forests. Machine learning, 45*, 5–32.

Centers for Disease Control and Prevention. Data Modernization Initiative. https://www.cdc.gov/surveillance/data-modernization/index.html (accessed November 2023).

Chan, V. C., Ross, G. B., Clouthier, A. L., Fischer, S. L., & Graham, R. B. (2022). The role of machine learning in the primary prevention of work-related musculoskeletal disorders: A scoping review. *Applied Ergonomics, 98*, Article 103574.

Cidell, J. (2010). Content clouds as exploratory qualitative data analysis. *Area, 42*(4), 514–523.

DePaolo, C. A., & Wilkinson, K. (2014). Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. *TechTrends, 58*(3), 38–44.

Friedman, J., Beletsky, L., & Schriger, D. L. (2021). Overdose-related cardiac arrests observed by emergency medical services during the US COVID-19 epidemic. *JAMA Psychiatry, 78*(5), 562–564.

Garg, R., Richards, C. T., Naidech, A., & Prabhakaran, S. (2019). Abstract tp296: Predicting Cincinnati prehospital stroke scale components in emergency medical services patient care reports using natural language processing and machine learning. *Stroke, 50*(Suppl_1), Article ATP296-ATP296.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60–76.

Haas, E. J., Furek, A., Casey, M., Yoon, K. N., & Moore, S. M. (2021). Applying the Social Vulnerability Index as a leading indicator to protect fire-based emergency medical service responders’ health. *International Journal of Environmental Research and Public Health, 18*(15), 8049.

Haas, E. J., Yoon, K. N., Furek, A., Casey, M., & Moore, S. M. (2023). The role of emergency incident type in identifying first responders’ health exposure risks. *Journal of Safety Science and Resilience, 4*(2), 167–173.

Henderson, S. (2009). Operations research tools for addressing current challenges in emergency medical services. In J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, & J. C. Smith (Eds.), *Wiley encyclopedia of operations research and management science* (pp. 1–16). New York: Wiley.

Hoff, C., Combs-Black, K., Sorek, J. D., Elsenboss, C., Robinson, M. M., & Robison, B. (2023). Public health emergency preparedness and response after COVID-19. *Health Security, 21*(S1), S72–S78.

International Public Safety Data Institute [IPSDI], https://i-psdi.org/, last accessed: June 2023.

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Prentice Hall.

Lasher, L., Jason Rhodes MPA, A. C., & Viner-Brown, S. (2019). Identification and description of non-fatal opioid overdoses using Rhode Island EMS data, 2016–2018. Rhode Island Medical Journal, 102(2), 41–45.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT Press, Cambridge, MA.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551.

National Fire Data Center (2015). U.S. Fire Administration. National Fire Incident Reporting System. VERSION 5.0 DESIGN DOCUMENTATION. Specification Release 2015.1, January 2015. FEMA. Available at: https://www.usfa.fema.gov/downloads/pdf/nfirs/nfirs_spec_2015.pdf.

National Fire Data Center (2017). U.S. Fire Administration. Review and Assessment of Data Quality in the National Fire Incident Reporting System. FEMA. Available at: https://www.usfa.fema.gov/downloads/pdf/publications/nfirs_data_quality_report.pdf.

Pitt, I.L. (2022). The system-wide effects of dispatch, response and operational performance on emergency medical services during COVID-19. Humanities and Social Sciences Communications, 9, 412. https://doi.org/10.1057/s41599-022-01405-z.

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology, 2*(1), 37–63.

Pramanik, B., & Modak, S. (2024). A Machine Learning Approach to Identifying Employees at Risk of Workplace Injury. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-6). IEEE.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, 3(22), 41-46.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: When is it useful? *Journal of Information Science, 34*(1), 15–29.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Suen, C. Y. (1979). N-gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 164–172.

U.S. Fire Administration (2023). U.S. Fire Administration Announces Effort to Launch New Fire Information and Analytics Platform. Published May 9, 2023 at: https://www.usfa.fema.gov/about/media-releases/2023-05-09-usfa-announces-new-fire-information-platform.html.

Valecha, R., Sharman, R., Rao, H. R., & Upadhyaya, S. (2013). A dispatch-mediated communication model for emergency response systems. *ACM Transactions on Management Information Systems (TMIS), 4*(1), 1–25.

Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., & McKeon, M. (2007). Manyeyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1121–1128.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS), 22*(2), 179–214.

N. Katherine Yoon, PhD: Oak Ridge Institute for Science and Education Fellow, National Personal Protective Technology Laboratory (NPPTL), National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC), 626 Cochrans Mill Road, Pittsburgh, PA 15236. Dr. Nami Katherine Yoon is an interdisciplinary research professional with Ph.D. in public health policy. Since 2020, she has been serving as an Oak Ridge Institute for Science and Education (ORISE) research fellow at CDC/NIOSH/NPPTL and her research focuses on health and safety of workers and personal protective equipment (PPE). Her research encompasses policy research and evaluation as well as data analysis employing diverse research design and methods both quantitative and qualitative. She received her Ph.D. from the University of Pittsburgh with focus on the approval process for medical products and post-market regulation. She has 15+ years of experience with policy analysis, policy research, project management, and data analysis in the private and public sector. Katherine has received her Master's degree in Public Policy from Harvard University. She also holds Bachelor of Engineering in Computer Engineering from Kyunghee University in South Korea.

Tyler D. Quinn, PhD: Assistant Professor, Department of Epidemiology and Biostatistics, School of Public Health, West Virginia University, PO Box 9190, Morgantown WV 26506-9190. Tyler Quinn is an assistant professor at the West Virginia University School of Public Health. He received a PhD in exercise physiology from the University of Pittsburgh and completed his pre- and post-doctoral training at the Centers for Disease Control and Prevention's National Institute for Occupational Safety and Health. His research portfolio focuses broadly on examining the impacts of occupational demands including physical activity, psychosocial stress, and environmental conditions on worker cardiometabolic health. His most recent research is examining the explanatory mechanisms to the paradoxical adverse associations between occupational physical activity and cardiovascular health.

Alexa Furek, MPH: Former CDC/NIOSH/NPPTL, Currently residing in Austin, TX. Alexa Furek, M.P.H., joined NIOSH's NPPTL as an Oak Ridge for Science and Education (ORISE) research fellow. She holds an M.P.H. in Behavioral and Community Health Sciences and a B.S. in Psychology from the University of Pittsburgh. While at NIOSH, Furek managed data for nationwide hospital PPE inventory monitoring systems as well as firefighter exposure data to monitor exposure characteristics and PPE usage.

Nora Y. Payne, PhD: Research Statistician, National Personal Protective Technology Laboratory (NPPTL), National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC), 626 Cochrans Mill Road, Pittsburgh, PA 15236. Dr. Nora Payne is a Research Statistician with the National Personal Protective Technology Laboratory, advising researchers on study design and statistical methodology and additionally conducting research related to the analysis of large-scale occupational health and safety data. Nora received her Ph.D. in Statistics in 2022 from the University of Michigan.

Emily J. Haas, PhD: Associate Director for Science, Division of Safety Research, National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC), 626 Cochrans Mill Road, Pittsburgh, PA 15236. Dr. Haas is the Associate Director for Science at the Division of Safety Research within the National Institute for Occupational Safety and Health. In her current role, Dr. Haas provides scientific, strategic, and operational leadership to support evidence-based research and practices. Her research involves developing, implementing, and evaluating organizational interventions with an emphasis on improving worker and management engagement in response to new technologies. Dr. Haas received her Ph.D. from Purdue University in Health Communication and her B.A./M.A. from the University of Dayton.