American Journal of
Preventive Medicine

RESEARCH ARTICLE

Characterizing Female Firearm Suicide Circumstances: A Natural Language Processing and Machine Learning Approach

Evan V. Goldstein, PhD, MPP,[1] Stephen J. Mooney, PhD,[2,3] Julian Takagi-Stewart, BS,[3,4,5] Brianna F. Agnew, MPH,[6] Erin R. Morgan, PhD,[2,3] Miriam J. Haviland, PhD, MSPH,[3] Weipeng Zhou, BS,[7] Laura C. Prater, PhD, MPH, MHA[3,8]

Introduction: Since 2005, female firearm suicide rates increased by 34%, outpacing the rise in male firearm suicide rates over the same period. The objective of this study was to develop and evaluate a natural language processing pipeline to identify a select set of common and important circumstances preceding female firearm suicide from coroner/medical examiner and law enforcement narratives.

Methods: Unstructured information from coroner/medical examiner and law enforcement narratives were manually coded for 1,462 randomly selected cases from the National Violent Death Reporting System. Decedents were included from 40 states and Puerto Rico from 2014 to 2018. Naive Bayes, Random Forest, Support Vector Machine, and Gradient Boosting classifier models were tuned using 5-fold cross-validation. Model performance was assessed using sensitivity, specificity, positive predictive value, F1, and other metrics. Analyses were conducted from February to November 2022.

Results: The natural language processing pipeline performed well in identifying recent interpersonal disputes, problems with intimate partners, acute/chronic pain, and intimate partners and immediate family at the scene. For example, the Support Vector Machine model had a mean of 98.1% specificity and 90.5% positive predictive value in classifying a recent interpersonal dispute before suicide. The Gradient Boosting model had a mean of 98.7% specificity and 93.2% positive predictive value in classifying a recent interpersonal dispute before suicide.

Conclusions: This study developed a natural language processing pipeline to classify 5 female firearm suicide antecedents using narrative reports from the National Violent Death Reporting System, which may improve the examination of these circumstances. Practitioners and researchers should weigh the efficiency of natural language processing pipeline development against conventional text mining and manual review.
Am J Prev Med 2023;65(2):278−285.

From the [1]Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, Utah; [2]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington; [3]Harborview Injury Prevention & Research Center, Harborview Medical Center, University of Washington, Seattle, Washington; [4]College of Medicine, Drexel University, Philadelphia, Pennsylvania; [5]Department of Anesthesiology & Pain Medicine, School of Medicine, University of Washington, Seattle, Washington; [6]School of Nursing and Health Professions, University of San Francisco, San Francisco, California; [7]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, Washington; and [8]Department of Psychiatry and Behavioral Sciences, School of Medicine, University of Washington, Seattle, Washington

Address correspondence to: Evan V. Goldstein, PhD, MPP, Department of Population Health Sciences, Spencer Fox Eccles School of Medicine, University of Utah, Williams Building Room 1N502, Salt Lake City UT 84108. E-mail: evan.goldstein@hsc.utah.edu.

0749-3797/$36.00
https://doi.org/10.1016/j.amepre.2023.01.030

INTRODUCTION

Suicide was the 12th leading cause of death for Americans across the lifespan in 2020.[1] Risk factors for suicide may include having had a previous attempt, mental illness, substance use disorder, stressful life events, or exposure to violence over the life course.[1–6] Having a firearm in the home is a strong independent risk factor for suicide; firearms account for more than half of all U.S. suicide deaths.[1,2,7] Stressful life events, often acute and episodic in nature, are more common among those who use firearms than among those who experience other methods of suicide death.[8]

Suicide attempts among females are often less lethal than among males, likely owing to mechanism choice. Females are less likely than males to use firearms, which are the most fatal suicide method.[1] Since 2005, however, female firearm suicide rates increased by 34%, outpacing the rise in male firearm suicide rates over the same period. Firearm use as a mechanism has increased at a faster rate among subgroups of females, including female veterans relative to nonveterans.[9,10] Little is known about circumstances or physical and mental health concerns preceding the death of females by firearm suicide.

The limited investigation into circumstances around female firearm suicide is in part attributable to limited data. The Centers for Disease Control and Prevention's (CDC) National Violent Death Reporting System (NVDRS)[11] is a primary source of detailed information on violent deaths across participating states. NVDRS data have been used extensively to study circumstances preceding suicides; however, studies have been limited in scope owing to the burden of manually coding circumstances extracted from narrative information provided by coroners/medical examiners (CMEs) and law enforcement (LE). Natural language processing (NLP), which uses algorithms to identify and extract themes from text, holds promise for efficiently identifying circumstances in NVDRS narratives, including new antecedents relevant to female firearm suicide deaths. This study had 2 objectives: first, to develop and share a practical NLP approach that could be helpful to NVDRS data users, and second, to address the gap in the empirical examination of circumstances preceding female firearm suicide.

METHODS

Study Sample

Researchers requested and received deidentified data from the NVDRS Restricted Access Database[12] on all female firearm suicides from 2014 to 2018 through a review and approval process managed by CDC. The NVDRS program links data from vital records, CME, and LE agencies to compile comprehensive incident-level data on violent deaths from all 50 states, the District of Columbia, and Puerto Rico. The system contains hundreds of elements, including demographic, injury and death, circumstance, weapon, and toxicology information (coded by trained abstractors) and unstructured CME and LE reports summarized from interviews of persons close to each death. These records are then linked to a variety of optional secondary data sources, including Supplementary Homicide Reports, National Incident-Based Reporting System, hospital data, and court records.[13]

The study team randomly selected 1,462 female firearm suicide cases for manual coding, which were representative of all female firearm suicide cases recorded in NVDRS from 2014 to 2018. Not all states reported to NVDRS at this time. Appendix Table A1 (available online) describes the states and years represented in the analytic sample. This study did not require approval from the University of Washington IRB because using deidentified data on deceased individuals was not considered human subjects research.

Measures

Because the etiology of female firearm suicide is understudied, a predetermined list of circumstances was not used. The study team reviewed a random sample of the CME/LE narratives from our study population,[14] noting recurrent characteristics preceding firearm suicides. Specifically, following best practices and previous literature on generating a data set for NLP model training,[15,16] 3 researchers (LCP, JTS, MH) each read 100 randomly selected CME and LE narratives and developed independent lists of key themes. The full research team then consolidated these lists and reached a consensus on which to include in the NLP pipeline.

Two researchers (JTS and BA) then reviewed all 1,462 female firearm suicide incidents in the analytic sample and coded occurrences of the preceding circumstance labels. For each female firearm suicide decedent, a preceding circumstance label equaled 1 if it was present in the CME or LE narratives and 0 if it was not present. The narrative CME and LE reports were labeled and organized using Dedoose.[17] Researchers (JTS, BA) double coded 100 randomly selected narratives to ensure consistency in the application of codes and achieved agreement in all cases. On coding beyond the initial 100, labelers met weekly to discuss each instance of coding and ensure agreement. Where there was uncertainty, a third researcher (LCP) mediated and, when necessary, presented the questions to the full research team, facilitating discussion until agreement was achieved.

Of the coded labels, 5 occurred in more than 5% of all manually reviewed cases (Table 1[18]). For developing NLP pipelines, the study team focused on the 5 prevalent labels because of a concern that label imbalance and inadequate sample size would compromise the ability to develop reliable algorithms on the rarer labels. Appendix Table A2 (available online) lists the labels identified but not modeled with NLP.

Statistical Analyses

Conventional NLP techniques—including Snowball stemming, lemmatization, and Term Frequency Inverse Document Frequency weighting—were used to preprocess the free-text narrative reports for further analysis. All NLP preprocessing was conducted in a Python coding environment using the Beautiful Soup, Natural Language ToolKit, and Scikit-learn libraries.[19–21] Details about the preprocessing procedures are available in the Appendix

Table 1. Definitions and Counts of Manually Coded Preceding Circumstance Labels Applied to the Training Data Set (N=1,462)

Label	Definition	Count (%)
Intimate partner at the scene	The decedent's current spouse or partner was noted to be at the scene of the suicide.	425 (29.1)
Problems with intimate partner(s)	The decedent was noted to be experiencing problems with the most recent romantic relationship, such as an affair (either victim or partner), break up, divorce, or mention of a troubled relationship.	318 (21.8)
Recent interpersonal dispute	The decedent was noted to be involved in an interpersonal dispute including arguments and confrontational fallouts.	304 (20.8)
Immediate family at the scene	The decedent's child, parents, relatives, or siblings were noted to be at the scene of the suicide.	301 (20.6)
Acute/chronic pain	The decedent was noted to be experiencing acute or chronic pain as an explicit mention of physical pain or a noted diagnosis of a condition that is associated with chronic pain.[18]	243 (16.7)

(available online). Notably, other researchers have recently applied similar procedures to NVDRS data in attempts to overcome large-scale text classification problems and other suicide-related classification problems.[22,23]

Once all 1,462 labeled narratives were converted to numerical vector representations, the study team developed supervised machine learning (ML) algorithms to identify the occurrence of each of the 5 common labels. Specifically, Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB) classifier models were trained. These 4 approaches were chosen for their efficiency and potential ease of use by different stakeholders, ability to handle a large number of features, adaptability to nonlinearities in the data or nonlinear boundaries between classes, and robustness to outliers.[24−26]

To tune the models and test performance, the labeled analytic sample was randomly split into a training set (80%) and a test set (20%) for each label. On the training set, 5-fold cross-validation[27] was used to tune the model's hyperparameters to maximize F1 (the harmonic mean of precision positive predictive value [PPV] and sensitivity). The models were tuned for each label separately because a narrative could include more than 1 label. The Appendix (available online) provides additional information about the hyperparameters tuned through this process.

Final performance metrics were calculated by applying each model to a held-out test set. Model performance was evaluated for each preceding circumstance label by calculating sensitivity, specificity, PPV, F1, accuracy, and area under the receiver operating characteristic curve (AUC-ROC) for 5 random data split and model seeds. The Appendix (available online) describes these performance metrics. After reviewing the ML model performance, the research team manually reviewed classification results from the best-performing ML models to better contextualize errors in classification. Appendix Figure A1 (available online) summarizes the process for sampling and applying labels to the data set, preprocessing the data in the NLP pipeline, and classification modeling.

RESULTS

Table 2 describes the analytic sample of 1,462 female firearm suicides occurring from 2014 to 2018. The mean age of female firearm suicide decedents was 47.0 years. Most decedents were White (90.5%); 96.4% of the decedents were non-Hispanic or of unknown ethnicity. About 40.7% of decedents were married or in civil unions or domestic partnerships, and 87.4% of decedents had high school or greater educational attainment. Nearly half (43.2%) of decedents had a depression-related diagnosis, but only 29.3% of decedents were undergoing mental health treatment at the time of death.

Table 1,[18] summarizes the frequency of each manually coded preceding circumstance label included in the NLP pipeline. The prevalence of reported preceding circumstances ranged from 16.7% (for acute or chronic pain issues) to 29.1% (for intimate partner at the scene). Examples of text fragments that indicate a positive code for each label are presented in Table 3. For example, instances of *tired of her marriage* and *relationship issues* indicated an intimate partner problem. Text fragments such as *constant pain*, *lot of pain*, or specific conditions indicated acute/chronic pain.

Table 4 describes the classification performance metrics from the NB, RF, SVM, and GB models on the held-out test sets for 5 random data split and model seeds. Model performance varied depending on the label. This section reports results on 3 of the labels (recent interpersonal dispute, intimate partner problem, acute/chronic pain), leaving out the 2 labels describing who was at the scene after the death. The SVM model had a mean of 98.1% specificity, 70.5% sensitivity, 79.1 F1 value, 90.5% PPV, and 84.3 AUC-ROC in classifying whether decedents had a recent interpersonal dispute before suicide. In comparison, the GB model had a mean of 98.7% specificity, 67.7% sensitivity, 78.4 F1 value, 93.2% PPV, and 83.2 AUC-ROC in classifying whether decedents had a recent interpersonal dispute before suicide. The RF model had a mean of 99.4% specificity, 60.5% sensitivity, 74.0 F1 value, 96.1% PPV, and 79.9 AUC-ROC. The NB model had a mean of 80.4% specificity, 70.1% sensitivity, 56.6 F1 value, 47.5% PPV, and 75.2 AUC-ROC.

Table 2. Demographics and Characteristics of Female Firearm Suicide Decedents in the Analytic Sample (n=1,462), 2014–2018

Variables	Mean/n	SD/%
Age, mean (SD)	47.0	6.9
Race, n (%)		
White	1,323	90.5
Black or African American	79	5.4
Asian	26	1.8
American Indian/Alaska Native	15	1.0
All other	19	1.3
Ethnicity, n (%)		
Hispanic	52	3.6
Non-Hispanic or unknown	1,410	96.4
Marital status, n (%)		
Divorced, separated	337	23.1
Married/civil union/domestic partnership	596	40.7
Single	405	27.7
Widowed	120	8.2
Unknown/missing	4	0.3
Ever served in the military, n (%)	58	4.0
Education Level, n (%)		
High school or greater	1,330	87.4
Less than high school	132	9.0
Unknown	52	3.6
Chronic alcohol abuse, n (%)	209	14.3
Other substance abuse, n (%)	198	13.5
Current mental health problem, n (%)	793	54.2
Diagnosis,[1] n (%)		
Depression	631	43.2
Anxiety	160	11.0
Bipolar	132	9.0
Eating disorder	6	0.4
Obsessive-compulsive disorder	1	0.1
Post-traumatic stress disorder	27	1.9
Schizophrenia	11	0.8
Undergoing mental health treatment, n (%)	428	29.3
Suicidal thoughts/actions, n (%)		
History of suicidal thoughts	481	32.9
History of suicide attempts	338	23.1
Recent suicide of friend/family, n (%)	49	3.4
Other death of friend/family, n (%)	115	7.9
Job problem, n (%)	122	8.3
Financial problem, n (%)	120	8.2
Eviction/loss of home, n (%)	53	3.6
Recent criminal/legal problem, n (%)	46	3.2

The SVM model had a mean of 96.3% specificity, 54.5% sensitivity, 64.4 F1 value, 80.7% PPV, and 75.4 AUC-ROC in classifying whether decedents had a problem with an intimate partner before suicide. In comparison, the GB model had a mean of 96.5% specificity, 57.0% sensitivity, 67.0 F1 value, 81.5% PPV, and 76.8 AUC-ROC in classifying whether decedents had a problem with an intimate partner before suicide. The RF model had a mean of 98.2% specificity, 30.3% sensitivity, 44.2 F1 value, 82.6% PPV, and 64.2 AUC-ROC. The NB model had a mean of 70.5% specificity, 68.3% sensitivity, 50.6 F1 value, 40.2% PPV, and 69.4 AUC-ROC.

Table 3. Examples of Manually Coded Preceding Circumstance Labels Applied to the Training Data Set

Label	Coded examples
Problems with intimate partner(s)	Affair, restraining order out against her husband, separated husband, cheating, divorced, separated, relationship issues, broke up, marital issues, tired of her marriage
Recent interpersonal dispute	Argue, completely disowned, phone call was not going well, domestic disputes, became verbally aggressive, argument, walked off, not allowed to go, verbal altercation, screaming at him
Intimate partner at the scene	Found dead by her boyfriend, boyfriend returned to the living room, husband woke up, husband heard a loud crashing sound, girlfriend witnessed the victim put the gun to her head and pull the trigger
Immediate family at the scene	Family heard the gunshot, discovered by her mother, victim's sister heard, found by her son, found by her father, brother went to check on the victim, found death by her daughter, found by her parents
Acute/chronic pain	Migraine, spine fracture, osteoporosis, knee surgery, chronic pain, constant pain, osteoarthritis, fibromyalgia, arthritis, irritable bowel syndrome, multiple sclerosis, lot of pain

The SVM model had a mean of 98.7% specificity, 65.7% sensitivity, 76.7 F1 value, 92.3% PPV, and 82.2 AUC-ROC in classifying acute/chronic pain before suicide. The GB model had a mean of 98.4% specificity, 75.3% sensitivity, 81.9 F1 value, 89.9% PPV, and 86.8 AUC-ROC. The RF model had a mean of 99.3% specificity, 54.3% sensitivity, 68.5 F1 value, 94.5% PPV, and 76.8 AUC-ROC. The NB model had a mean of 90.4% specificity, 52.9% sensitivity, 52.3 F1 value, 53.8% PPV, and 71.6 AUC-ROC.

Table 4. Performance Metrics of the Naive Bayes, Random Forest, Support Vector Machine, and Gradient Boosting Models

Label	1, Naive Bayes		2, Random Forest		3, Support Vector Machine		4, Gradient Boosting	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Recent interpersonal dispute								
Sensitivity (%)	70.1	2.6	60.5	6.4	70.5	2.4	67.7	3.1
Specificity (%)	80.4	4.4	99.4	0.2	98.1	1.1	98.7	0.7
Positive predictive value (%)	47.5	2.9	96.1	1.2	90.5	5.3	93.2	3.7
F1	56.6	1.6	74.0	4.8	79.1	1.4	78.4	3.0
Accuracy (%)	78.4	3.3	92.0	1.0	92.5	0.4	92.4	1.2
AUC-ROC	75.2	1.4	79.9	3.1	84.3	0.8	83.2	1.8
Problems with intimate partner(s)								
Sensitivity (%)	68.3	7.5	30.3	3.9	54.5	12.5	57.0	6.2
Specificity (%)	70.5	3.6	98.2	0.7	96.3	1.2	96.5	1.6
Positive predictive value (%)	40.2	3.4	82.6	6.2	80.7	4.7	81.5	8.1
F1	50.6	4.6	44.2	3.9	64.4	10.1	67.0	6.4
Accuracy (%)	70.2	1.5	83.1	2.1	86.9	3.6	88.5	2.1
AUC-ROC	69.4	2.2	64.2	1.8	75.4	6.1	76.8	3.5
Acute/chronic pain								
Sensitivity (%)	52.9	10.8	54.3	10.6	65.7	2.1	75.3	3.9
Specificity (%)	90.4	2.5	99.3	0.3	98.7	0.5	98.4	0.6
Positive predictive value (%)	53.8	6.6	94.5	3.0	92.3	2.8	89.9	4.2
F1	52.3	4.4	68.5	8.1	76.7	1.1	81.9	3.2
Accuracy (%)	83.6	1.7	91.5	1.7	92.7	0.6	94.6	0.8
AUC-ROC	71.6	4.4	76.8	5.2	82.2	0.9	86.8	2.0
Intimate partner at the scene								
Sensitivity (%)	78.8	0.9	51.5	5.7	73.0	3.6	66.7	4.9
Specificity (%)	83.6	3.0	92.7	5.0	93.2	1.8	93.9	2.2
Positive predictive value (%)	66.7	3.4	79.6	7.6	81.4	5.0	79.7	3.1
F1	72.2	1.8	62.3	4.6	76.8	1.0	72.6	3.8
Accuracy (%)	82.2	2.0	82.6	2.1	87.3	1.2	86.8	4.1
AUC-ROC	81.2	1.2	73.1	2.7	83.1	1.2	80.3	3.2
Immediate family at the scene								
Sensitivity (%)	67.7	5.3	20.2	3.3	57.0	10.8	50.2	14.8
Specificity (%)	88.3	2.8	97.7	1.6	93.6	1.1	95.8	1.9
Positive predictive value (%)	59.8	6.5	71.8	13.4	69.0	2.4	74.2	10.4
F1	63.4	5.6	31.2	3.8	62.0	7.1	59.6	13.3
Accuracy (%)	84.2	2.5	81.9	2.4	86.1	2.3	86.6	5.1
AUC-ROC	78.0	3.0	58.9	1.5	75.3	5.1	73.0	8.1

Note: This table shows the ML model performance metrics following the NLP preprocessing steps described in Methods. The mean performance metrics of 5 random data split and model seeds are shown for each algorithm. Performance variability is shown using SDs. F1 denotes the harmonic mean of positive predictive value and sensitivity.
AUC-ROC, area under the receiver operating characteristic curve; ML, machine learning; NLP, natural language processing.

The performance metrics using token lemmatization were similar to the metrics described in Table 4 for the recent interpersonal dispute, problems with intimate partners, intimate partner at the scene, immediate family at the scene, and acute/chronic pain labels, depending on the label, model, and metric (Appendix Table A3, available online).

Samplings of classification errors from the best-performing SVM models for the 5 labels were manually reviewed. Incorrect classification examples were typically false negatives. For example, false negatives regarding whether decedents had a recent interpersonal dispute described the disputes using vague or indirect language or the passive voice. In comparison, correctly classified recent interpersonal disputes included explicit language about the conflict, for example, statements such as, "The victim had an argument with her ex-boyfriend that day." Similarly, vague or indirect language was also a challenge in identifying problems with intimate partners. Typical false negatives included statements such as, "The victim was reportedly having some issues with a current boyfriend." Correctly classified intimate partner conflicts used direct statements, such as, "The victim just went through a rough divorce." Other labels (intimate partner at scene, immediate family at the scene, and acute/chronic pain) had similar trends of vague language in cases of classification errors.

DISCUSSION

This study resulted in the development of a comprehensive and successful NLP pipeline for 5 common and important circumstantial variables preceding female firearm suicide death. Circumstances relevant to suicide prevention were identified, such as identifying intimate partner problems and acute/chronic pain, with relatively high PPV (>80.0%). These findings suggest that NLP is a promising approach to addressing a critical gap in the identification and examination of factors commonly contributing to female firearm suicide death. Findings from this study, the associated code, and coding documentation may be useful for public health surveillance and practice as well as for other researchers using NVDRS data.

Several other studies have used NLP approaches to classify on a single concept in NVDRS, with varying success.[22,23,28] This study complements this work by focusing on the unique antecedents of female firearm suicide. This study shows that using a broad list of labels appears to be most fruitful for common characteristics described by direct language in the NVDRS CME and LE narratives. The study team also found that using the Term Frequency Inverse Document Frequency weighting technique and both token stemming and lemmatization improved a model's performance compared with not using these techniques. However, extensive resources may be required to develop a general-purpose NLP tool. The process of manually reviewing the CME and LE narratives for a representative sample of 1,462 suicide decedents was resource intensive for the study team given a 1-year project timeline. To develop an NLP pipeline that performs generally well for common circumstances preceding firearm suicide deaths among female decedents and other populations requires a substantial investment in personnel, time, and expertise. Assuming that these resources are not commonly available to public health practitioners and researchers who wish to apply NLP algorithms to NVDRS, key decisions need to be made about whether to apply NLP or use a text mining plus manual review approach.[29,30]

Alternatively, weak forms of supervision or few-shot learning techniques may help to address limited resources. For weak supervision approaches, researchers could explore ways to programmatically automate the assignment of noisy labels to raw NVDRS narratives using rules, dictionaries, crowd-sourced annotation, and statistical modeling preceding ML model training.[31−33] Researchers could also test few-shot learning approaches using ML models to discriminate classes using only a small, finely tuned amount of training data and *learning to learn* similarities across the circumstance classes.[34] The impact of these additional approaches on model performance in comparison with that of full supervision using many manually labeled narratives is promising but remains unclear. Bidirectional long short-term memory networks have also shown promising results in detecting suicidal tendencies in different contexts by increasing the amount of information available to a neural network through 2 long short-term memory layers.[35]

In addition to identifying a pathway for an NLP pipeline, this study describes several important circumstances frequently preceding female firearm suicide. Females who died of firearm suicide often had an intimate partner or family member at the scene of the suicide, often present during the death or first on the scene after the death. Exposure to suicide among those close to the decedent causes far-reaching effects, including increased risk for mental illness and suicidal ideation.[36] Having problems with an intimate partner and having a recent interpersonal dispute were common preceding circumstances for female firearm suicides. Evidence clearly links intimate partner violence to suicidal thoughts and behaviors,[37,38] but this study shows that nonviolent forms of conflict may also be worthy of examination. Finally, about 20% of the sample had acute or chronic pain. Previous studies have estimated chronic pain to be closer to 10% among suicide decedents of both sexes using all mechanisms.[39] Moreover, pain may be undertreated or differentially treated among females relative to that among males.[40] More research on suicide and pain among females is necessary, particularly among those living in a home with a firearm.

Limitations

There are several limitations to the approach described in this paper. First, CME and LE narratives in NVDRS are limited to what is reported in next-of-kin interviews and what is determined relevant to the death by the coroner, medical examiner, and/or LE officer. Accordingly, the absence of a circumstance in the narrative does not mean it was not present for the decedent. Second, CDC abstractors summarize full CME/LE reports when writing the narrative summaries from which the NLP models were trained, potentially leaving out relevant details or, at worst, creating inconsistencies between the narrative and the full report. Third, racial/ethnic differences in the length of narrative and choice of language used in NVDRS narratives may create biases in NLP models,[41] implying that the use of NLP to automate identifying antecedents could further perpetuate disparities in information on suicide among Black or other racial/ethnic minority persons. The authors are examining model performance differences by different demographic and geographic characteristics in a follow-up study. Fourth, incorrect classification examples were typically false negatives. Applying the models developed during this study could help to classify interpersonal disputes, problems with intimate partners, acute/chronic pain, and intimate partners and immediate family at the scene before suicide death with high precision and specificity and reduce the time and resources otherwise needed to manually code these circumstances. However, as a trade off, users must be mindful that the models may underestimate the actual occurrence of certain circumstances in comparison with manual review owing to false-negative classification. False negatives were likely because true negative rates were high in this context where no label was present in >30% of the population. Finally, although this study's findings may help improve public health surveillance, the study only examined data from decedents only and should not be used to predict or prevent suicidal behavior, which was not the intent of the study.

CONCLUSIONS

This study shows the feasibility of developing an NLP pipeline to classify common antecedents of female firearm suicide using narratives from NVDRS. This study shows that SVM and GB models performed better than RF and NB models in classifying these common circumstances. However, practitioners and researchers considering developing such pipelines should carefully weigh the efficiency of NLP pipeline development against that of conventional text mining and manual review. With a significant investment in resources, a large enough training data set, and oversampling of underrepresented populations, researchers could achieve the goal of creating a widely usable NLP pipeline for NVDRS label classification. Future directions include investigating NLP model performance by vulnerable subgroups (e.g., on the basis of race/ethnicity, age, and rurality) to ensure that models perform equitably.

ACKNOWLEDGMENTS

The sponsor of this study had no role in the study design; collection, analysis, and interpretation of data; writing of the report; or the decision to submit this paper for publication.

This study was supported by a grant from the National Collaborative on Gun Violence Research.

No financial disclosures were reported by the authors of this paper.

CREDIT AUTHOR STATEMENT

Evan V. Goldstein: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. Stephen J. Mooney: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. Julian Takagi-Stewart: Formal analysis, Writing – original draft. Brianna F. Agnew: Formal analysis, Writing – original draft. Erin R. Morgan: Resources, Writing – review & editing. Miriam Haviland: Data curation, Funding acquisition, Resources, Writing – review & editing. Weipeng Zhou: Formal analysis, Methodology, Validation. Laura C. Prater: Conceptualization, Data curation, Funding acquisition, Formal analysis, Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing.

SUPPLEMENTAL MATERIAL

Supplemental materials associated with this article can be found in the online version at https://doi.org/10.1016/j.amepre.2023.01.030.

REFERENCES

1. Web-Based injury statistics query and reporting system (WISQARS). U.S.: Centers for Disease Control and Prevention; 2021. Updated December 2, 2021. Accessed February 16, 2023. https://www.cdc.gov/injury/wisqars/index.html.
2. Miller M, Hemenway D. The relationship between firearms and suicide: a review of the literature. *Aggress Violent Behav.* 1999;4(1):59–75. https://doi.org/10.1016/S1359-1789(97)00057-8.
3. Swanson JW, McGinty EE, Fazel S, Mays VM. Mental illness and reduction of gun violence and suicide: bringing epidemiologic research to policy. *Ann Epidemiol.* 2015;25(5):366–376. https://doi.org/10.1016/j.annepidem.2014.03.004.
4. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Results from the 2014 National Survey on Drug Use and Health: detailed tables. Rockville, MD: Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality; 2015. Accessed August 16, 2022 https://www.samhsa.gov/data/report/results-2014-national-survey-drug-use-and-h..
5. Anestis MD. Prior suicide attempts are less common in suicide decedents who died by firearms relative to those who died by other means. *J Affect Disord.* 2016;189(1):106–109. https://doi.org/10.1016/j.jad.2015.09.007.
6. Shenassa ED, Catlin SN, Buka SL. Lethality of firearms relative to other suicide methods: a population based study. *J Epidemiol Community Health.* 2003;57(2):120–124. https://doi.org/10.1136/jech.57.2.120.
7. Romero MP, Wintemute GJ. The epidemiology of firearm suicide in the United States. *J Urban Health.* 2002;79(1):39–48. https://doi.org/10.1093/jurban/79.1.39.
8. Kaplan MS, McFarland BH, Huguet N. Characteristics of adult male and female firearm suicide decedents: findings from the National Violent Death Reporting System. *Inj Prev.* 2009;15(5):322–327. https://doi.org/10.1136/ip.2008.021162.
9. Kaplan MS, McFarland BH, Huguet N. Firearm suicide among veterans in the general population: findings from the national violent death

reporting system. *J Trauma.* 2009;67(3):503–507. https://doi.org/10.1097/TA.0b013e3181b36521.
10. McCarten JM, Hoffmire CA, Bossarte RM. Changes in overall and firearm veteran suicide rates by gender, 2001−2010. *Am J Prev Med.* 2015;48(3):360–364. https://doi.org/10.1016/j.amepre.2014.10.013.
11. National Violent Death Reporting System (NVDRS). Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html. Updated September 28, 2021. Accessed February 1, 2022.
12. NVDRS data access. Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/datasources/nvdrs/dataaccess.html. Updated September 28, 2021. Accessed October 1, 2022.
13. National Violent Death Reporting System (NVDRS) coding manual. Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/pdf/nvdrs/nvdrsCodingManual.pdf. Updated April 22, 2021. Accessed October 1, 2022.
14. Davidson JE, Ye G, Parra MC, et al. Job-related problems prior to nurse suicide, 2003−2017: a mixed methods analysis using natural language processing and thematic analysis. *J Nurs Regul.* 2021;12(1):28–39. https://doi.org/10.1016/S2155-8256(21)00017-X.
15. Stemler S. An overview of content analysis. *Pract Assess Res Eval.* 2000;7(1):17. https://doi.org/10.7275/z6fm-2e34.
16. McDonald N, Schoenebeck S, Forte A. Reliability and inter-rater reliability in qualitative research. *Proc ACM Hum Comput Interact.* 2019;3(CSCW):1–23. https://doi.org/10.1145/3359174.
17. Dedoose Version 8.0.35. Dedoose. https://www.dedoose.com/. Accessed February 1, 2022.
18. Dworkin RH, Bruehl S, Fillingim RB, Loeser JD, Terman GW, Turk DC. Multidimensional diagnostic criteria for chronic pain: introduction to the ACTTION−American Pain Society Pain Taxonomy (AAPT). *J Pain.* 2016;17(9):T1–T9 suppl. https://doi.org/10.1016/j.jpain.2016.02.010.
19. Bird S, Loper E, Klein E. Natural Language ToolKit (NLTK) Book. Sebastopol, CA: O'Reilly Media, 2009.
20. Richardson L. Beautiful soup documentation. *Read the Docs.*2016. https://beautiful-soup-4.readthedocs.io/en/latest/index.html?highlight=Beautiful%20soup%20documentation.
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.
22. Mezuk B, Ko TM, Kalesnikava VA, Jurgens D. Suicide among older adults living in or transitioning to residential long-term care, 2003 to 2015. *JAMA Netw Open.* 2019;2(6):e195627. https://doi.org/10.1001/jamanetworkopen.2019.5627.
23. Ko TM, Kalesnikava VA, Jurgens D, Mezuk B. A data science approach to estimating the frequency of driving cessation associated suicide in the U.S.: evidence from the national violent death reporting system. *Front Public Health.* 2021;9(1):689967. https://doi.org/10.3389/fpubh.2021.689967.
24. Schonlau M, Zou RY. The random forest algorithm for statistical learning. *The Stata Journal.* 2020;20(1):3–29. https://doi.org/10.1177/1536867X20909688.
25. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
26. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning - With Applications in R.* Cham, Switzerland: Springer, 2013.
27. Schaffer C. Selecting a classification method by cross-validation. *Mach Learn.* 1993;13(1):135–143. https://doi.org/10.1007/BF00993106.
28. Davidson JE, Proudfoot J, Lee K, Zisook S. Nurse suicide in the United States: analysis of the Center for Disease Control 2014 National Violent Death Reporting System dataset. *Arch Psychiatr Nurs.* 2019;33(5):16–21. https://doi.org/10.1016/j.apnu.2019.04.006.
29. Lyons VH, Adhia A, Moe CA, et al. Risk factors for child death during an intimate partner homicide: a case-control study. *Child Maltreat.* 2021;26(4):356–362. https://doi.org/10.1177/1077559520983901.
30. Shawon RA, Adhia A, DeCou C, Rowhani-Rahbar A. Characteristics and patterns of older adult homicides in the United States. *Inj Epidemiol.* 2021;8(1):5. https://doi.org/10.1186/s40621-021-00299-w.
31. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. 2009;2:1003−1011. https://doi.org/10.3115/1690219.1690287.
32. Lison P, Barnes J, Hubin A, Touileb S. Named entity recognition without labelled data: a weak supervision approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. https://doi.org/10.18653/v1/2020.acl-main.139.
33. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings VLDB Endowment.* 2017;11(3):269–282. https://doi.org/10.14778/3157794.3157797.
34. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. ArXiv. Preprint. Online May 28, 2020. https://arxiv.org/abs/2005.14165. Accessed November 7, 2022.
35. Gunjan VK, Vijayalata Y, Valli S, Kumar S, MO Mohamed, Saravanan V. Machine learning and cloud-based knowledge graphs to recognize suicidal mental tendencies. *Comput Intell Neurosci.* 2022;2022:3604113. https://doi.org/10.1155/2022/3604113.
36. Cerel J, Maple M, van de Venne J, Moore M, Flaherty C, Brown M. Exposure to suicide in the community: prevalence and correlates in one U.S. State. *Public Health Rep*. 2016;131(1):100−107. https://doi.org/10.1177/003335491613100116.
37. Iovine-Wong PE, Nichols-Hadeed C, Thompson Stone J, et al. Intimate partner violence, suicide, and their overlapping risk in women veterans: a review of the literature. *Mil Med.* 2019;184(5−6):e201–e210. https://doi.org/10.1093/milmed/usy355.
38. Devries KM, Mak JY, Bacchus LJ, et al. Intimate partner violence and incident depressive symptoms and suicide attempts: a systematic review of longitudinal studies. *PLoS Med.* 2013;10(5):e1001439. https://doi.org/10.1371/journal.pmed.1001439.
39. Petrosky E, Harpaz R, Fowler KA, et al. Chronic pain among suicide decedents, 2003 to 2014: findings from the national violent death reporting system. *Ann Intern Med.* 2018;169(7):448–455. https://doi.org/10.7326/M18-0830.
40. Hirsh AT, Hollingshead NA, Matthias MS, Bair MJ, Kroenke K. The influence of patient sex, provider sex, and sexist attitudes on pain treatment decisions. *J Pain.* 2014;15(5):551–559. https://doi.org/10.1016/j.jpain.2014.02.003.
41. Rahman N, Mozer R, McHugh RK, Rockett IRH, Chow CM, Vaughan G. Using natural language processing to improve suicide classification requires consideration of race. *Suicide Life Threat Behav.* 2022;52(4):782–791. https://doi.org/10.1111/sltb.12862.