Journal of the American Medical Informatics Association, 29(3), 2022, 472–480
https://doi.org/10.1093/jamia/ocab261
Advance Access Publication Date: 13 December 2021
Research and Applications

Research and Applications

Detection of self-harm and suicidal ideation in emergency department triage notes

Vlada Rozova [1,2], Katrina Witt[3,4], Jo Robinson[3,4], Yan Li[2], and Karin Verspoor [1,2]

[1]School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia, [2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia, [3]Orygen, Melbourne, Victoria, Australia, and [4]Centre for Youth Mental Health, The University of Melbourne, Melbourne, Victoria, Australia

Corresponding Author: Vlada Rozova, PhD, School of Computing Technologies, RMIT University, 124 La Trobe Street, Melbourne, VIC 3000, Australia; vlada.rozova@rmit.edu.au

Received 19 August 2021; Revised 30 September 2021; Editorial Decision 26 October 2021; Accepted 11 November 2021

ABSTRACT

Objective: Accurate identification of self-harm presentations to Emergency Departments (ED) can lead to more timely mental health support, aid in understanding the burden of suicidal intent in a population, and support impact evaluation of public health initiatives related to suicide prevention. Given lack of manual self-harm reporting in ED, we aim to develop an automated system for the detection of self-harm presentations directly from ED triage notes.

Materials and methods: We frame this as supervised classification using natural language processing (NLP), utilizing a large data set of 477 627 free-text triage notes from ED presentations in 2012–2018 to The Royal Melbourne Hospital, Australia. The data were highly imbalanced, with only 1.4% of triage notes relating to self-harm. We explored various preprocessing techniques, including spelling correction, negation detection, bigram replacement, and clinical concept recognition, and several machine learning methods.

Results: Our results show that machine learning methods dramatically outperform keyword-based methods. We achieved the best results with a calibrated Gradient Boosting model, showing 90% Precision and 90% Recall (PR-AUC 0.87) on blind test data. Prospective validation of the model achieves similar results (88% Precision; 89% Recall).

Discussion: ED notes are noisy texts, and simple token-based models work best. Negation detection and concept recognition did not change the results while bigram replacement significantly impaired model performance.

Conclusion: This first NLP-based classifier for self-harm in ED notes has practical value for identifying patients who would benefit from mental health follow-up in ED, and for supporting surveillance of self-harm and suicide prevention efforts in the population.

Key words: self-harm, natural language processing, machine learning, emergency department, suicidal ideation

INTRODUCTION

Globally, suicide is the second-leading cause of death for 15- to 19-year-olds,[1] and the leading cause of death for Australians aged between 15 and 44 years.[2] In Australia, expenditure on suicide prevention measures has increased from $1.9 million in 1995–1996 to $49.1 million in 2015–2016.[3] Despite such efforts, a significant reduction in suicide rates remains elusive: suicide currently claims nine Australian lives each day.[2]

For every death by suicide, 36 males and 280 young women are hospitalized following an episode of self-harm,[4] which we define

consistently with international conventions as any intentional act of self-injury, self-poisoning, or intentional drug overdose irrespective of the type of motivation or degree of suicidal intent.[5] Self-harm, and particularly frequently repeated self-harm, is also one of the strongest risk factors for suicide.[6] As such, hospital-presenting self-harm represents a useful proxy for evaluation of suicide prevention measures.

As a consequence, the World Health Organization (WHO) recommended a national or subnational surveillance system for all member states. Surveillance systems have therefore been established around the world,[7,8] including Australia.[9,10] However, these are not without their limitations.[11] The oldest continually operational system, the Multicentre Study of Self-Harm in England,[12] relies on manual identification of self-harm-related cases. Given this, there can be substantial lags between data collection and dissemination of outcomes, limiting the ability to monitor data in real time.[11]

Within Australia, the Hunter Area Toxicology Service (HATS) maintains a dedicated system to monitor self-harm presentations to clinical services. This system monitors intentional drug overdoses and self-poisoning presentations,[9] again based on manual case identification. Moreover, while self-poisoning is the most prevalent method of hospital-presenting self-harm, within the community certain methods of self-injury are more common.[13] Thus, this system is inadequate to monitor all forms of self-harm. The Victorian Injury Surveillance Unit (VISU) maintains the Victorian Emergency Minimum Dataset through all public EDs in Victoria. The data set compiles all injury-related presentations to EDs, including self-harm. However, case classification is dependent upon diagnosis (ICD-10) and human intent fields. Sensitivity of ICD-10 codes in identifying suicide ranges from 13.5% to 65%, and accuracy may be compromised by incomplete or inaccurate entry into the system.[14] As such, case classification that relies solely on ICD-10 codes is inaccurate, subject to bias in interpretation, and often underestimates the number of hospital-presenting self-harm cases.

In this work, we describe the development of an automated tool for identification of self-harm cases through natural language processing of ED triage notes. The method achieves strong performance, despite the tiny proportion of positive cases in the data set. It has the potential to be deployed to provide timely and actionable monitoring of self-harm presentations in the ED.

BACKGROUND AND SIGNIFICANCE

Electronic health records (EHRs) contain rich information and have been increasingly used to detect various medical conditions. Machine learning (ML) and natural language processing (NLP) offer wide-ranging solutions to retrieve and classify data from EHRs. In relation to suicide, researchers have attempted to leverage data from structured clinical fields in EHRs to predict suicide death and understand factors contributing to suicide risk. For example, a study by Choi et al[15] included over 800 000 people of which 2500 died by suicide. Structured data including sex, age, type of insurance, household income, disability, and medical records corresponding to eight ICD-10 codes were analyzed to predict the 10-year probability of suicide and identify risk factors.

The use of NLP for mental health applications is relatively understudied, as compared to other clinical applications,[16] but with significant potential to support monitoring, classification, and prediction of mental health illnesses.[17] NLP has seen limited use specifically in the context of suicidal ideation or direct self-harm. Recently, Carson et al[18] analyzed clinical notes collected from a small sample ($n = 73$) of psychiatrically hospitalized adolescents prior to their admission to detect both suicide attempts and suicidal ideation. Free text was linked to the Unified Medical Language System (UMLS) and converted into the UMLS concepts later used as features in a Random Forest model. Similarly, Fernandes et al[19] developed a hybrid model using a support vector machine (SVM) classifier applied to a bag-of-words followed by a set of heuristic rules aimed to reduce the number of false positives to detect suicide attempts in a psychiatric database. Further, neural networks have been reported to perform well at detecting intentional self-harm and predicting future attempts using a combination of progress notes, plan of care, and ED notes.[20] To our knowledge, our study is the first to focus exclusively on ED notes to identify self-harm.

ED triage notes are rapidly written short texts, with characteristics of ungrammaticality, spelling mistakes, and heavy use of clinical concepts and abbreviations, posing a challenge to traditional NLP methods. A multistage preprocessing including spelling correction and synonym integration is often required to harmonize nursing triage notes prior to further analysis. A recent study explored whether triage notes combined with structured information such as vitals and demographics can be used to predict sepsis from ED presentations.[21] The authors performed negation and bigram detection, common NLP techniques, and evaluated two different methods of text representation.

Another approach to analyzing structured and unstructured data collected from ED visits was reported by Gligorijevic et al.[22] To predict the number of resources an ED patient would require, the authors proposed a deep learning model based on the word attention mechanism to remedy the noisy text data in nursing notes.

We report on the development of a classification tool for detecting emergency department presentations related to self-harm. We compare several modeling approaches including keyword search, traditional machine learning (such as Logistic Regression), and deep learning. We also experiment with various preprocessing techniques to tackle the challenges associated with triage notes. Finally, we perform prospective validation of the model and examine the ability of the model to discern between self-harm (SH) and suicidal ideation (SI) cases.////

MATERIALS AND METHODS

Data collection

We collect a large data set of Emergency Department (ED) records from ED presentations at the Royal Melbourne Hospital (RMH) located in Melbourne, Australia. All ED presentations during the years 2012–2018 were extracted from the hospital's patient management systems, amounting to 477 627 rows of data.

Data annotation

For the purpose of this study, we considered only the textual component of each patient encounter, which records a brief note describing the reason for the presentation to an ED as per nursing assessment. Any additional structured data recorded for each encounter were excluded and the annotators were blinded to these during the annotation process. The average length of nursing triage notes was 127 characters; presentations with notes shorter than 30 characters were excluded from the data sets as they did not provide enough information. Notes from 2012 through to 2017 were reviewed by a trained postdoctoral level researcher with expertise in suicide prevention (author KW and a second annotator) and either labeled as a case of

self-harm (SH) or left unlabeled. The interannotator agreement calculated as Kohen's Cappa score was 0.91. Therefore, for our analysis, we considered two categories: triage notes with a positive SH annotation and controls. The data were highly imbalanced, with only 1.4% of all triage notes relating to SH.

Additional data collected during 2018 were used as a hold-out set for assessing the prospective application of the model. This data was annotated later for both SH and SI, identifying 1.6% and 2% of SH and SI presentations, respectively (Figure 1). It is worth mentioning that SH and SI cases are considered mutually exclusive; an SH presentation is more serious than an SI presentation. SH indicates the person has actually done something to hurt themselves, whereas SI indicates that a person is thinking about suicide, and may have made a plan (up to and including testing out a suicide method), but has not actually hurt themselves yet. We will examine these SI cases for confusion with SH. The lexical diversity of the development and hold-out sets calculated as the number of unique tokens divided by the total number of tokens is reported in Supplementary Table S1.

Preprocessing and tokenization

The first step of our text processing pipeline is tokenization which aims to transform a text string into a sequence of separated word-like units or *tokens*. This can usually be done by splitting a sentence by whitespace and isolating punctuation.

However, rapidly written triage notes contain a wide range of abbreviations that often involve punctuation and thus interact with typical tokenization rules. Examples include using "l)" for "left," "o/d" for "overdose," "++ve" for "positive," "r/ship" and "relationship," etc. This poses a significant challenge to performing accurate tokenization. We applied custom preprocessing to handle punctuation and expand such abbreviations.

We then adapted the scispaCy tokenizer[23] to our data, to split sentences by whitespace, separates commas and brackets while preserving complex numbers (eg, time, dates, vitals), and handling domain-relevant phrases like "o/d," "c/o" (complaining of) as a single token.

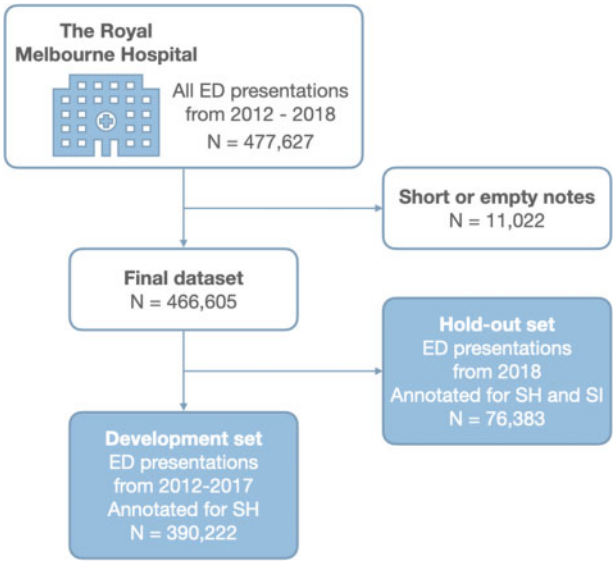

Figure 1. Summary of the RMH triage notes data set. Short and empty notes were excluded, and the resulting final annotated data set was split into development and hold-out sets based on the date of presentation. ED: emergency department; SH: self-harm; SI: suicidal ideation.

Domain-specific vocabulary

Domain-sensitive processing is needed due to the heavy use of specialized terms in ED texts. Therefore, we generated a domain-specific vocabulary based on the RMH ED triage corpus in the following way. First, we retrieved the vocabulary from a model trained on MIMIC-III free-text EHRs developed to enable clinical named-entity recognition.[24,25] Next, we removed all the words that did not appear in our data set. Finally, we manually added names of the local mental health organizations ("ECATT," "SAAPU," and "Orygen") and a list of common medication names for any indication, including generic names, brand names, and slang.

For each entry in the vocabulary, we counted its occurrence in the RMH ED triage corpus. This allowed us to generate a dictionary of 36 506 correctly spelled words and their frequencies and filter out terms that are irrelevant and likely to be noise in the context of the ED triage modeling. We refer to this as a word frequency list.

Text normalization and spelling correction

In the second step of our pipeline, we aim to reduce some of the variation in triage notes by correcting misspelled words and unifying synonymous terms, including drug names and medical abbreviations.

Composite tokens such as "warm/pink/dry" were further divided into parts in cases where each component of the token existed independently in our vocabulary.

For spelling correction, we used a Python implementation of a spell-checking algorithm provided in the `pyspellchecker` package.[26] When encountering an out-of-vocabulary token, the algorithm uses the Levenshtein (edit) distance to compare all possible permutations of the word within a specified distance to known words in a word frequency list and suggests the most probable correction. We identified and corrected 60 561 unique misspellings in our RMH ED triage corpus, thereby considerably reducing the variability of the data.

Additionally, slang names for common drugs were replaced by their generic names (for example, "Xanies" was changed to "Alprazolam" and "Fent" to "Fentanyl") to allow linking to a knowledge database such as Medical Subject Headings. Taken together, these steps reduced the dimensionality of the vocabulary from 100 328 to 43 887 unique tokens. For the impact of the preprocessing and spelling correction steps see Supplementary Table S2.

Model development

Following tokenization and spelling correction, we parsed the cleaned triage notes using a publicly available scispaCy pipeline trained on biomedical data and removed stop words, standalone punctuation, and numbers. The result of this filtering was used as input for further transformations and ML models (see Supplementary materials).

Performance evaluation

The RMH triage note development set was randomly shuffled and split into training and test sets in 80:20 proportion. Each model was evaluated by running 10-fold cross-validation on the training set to estimate its ability to generalize on unseen data. All splits/folds were performed in a stratified fashion to preserve the distribution of class

labels in every partition, particularly important given the very low frequency of SH instances in the data.

We computed the area under the Precision-Recall curve (PR AUC) to compare various models. The final model was calibrated, and hyper-parameters were tuned using cross-validation. For both classes, we calculated the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). We report metrics averaged across classes (macro averages) including Precision, Recall, and F1 score, the harmonic mean of Precision and Recall. We compared different models under cross-validation and evaluated the best model only on the held-out test set to assess generalization performance. Finally, we employed the Local Interpretable Model-Agnostic Explanations (LIME) algorithm provided as a Python package to interpret model predictions and evaluate the effect of distinct features.[27]

The results of model development and validation are reported in accordance with the Transparent Reporting of multivariate prediction models for Individual Prognosis Or Diagnosis (TRIPOD) Statement.[28]

RESULTS

Development of the classification tool

Presented below are the results of the development of a binary classifier for SH detection.

Keyword classifier

Below we provide the list of the identified important top 20 unigrams and bigrams (keywords) selected for use in our simple lookup classifier, based on association to the SH class:

Unigrams:.

> *"intent," "ingested," "diazepam," "polypharmacy," "suicidal," "x," "self," "intentional," "od," "mg," "tablets," "razor," "depression," "temazepam," "attempt," "overdose," "harm," "superficial," "seroquel," "inflicted."*

Bigrams:.

> *"polypharmacy od," "mg diazepam," "inflicted lac," "self harm," "intentional od," "taken x," "od x," "suicidal intent," "self inflicted," "mg x," "superficial lacs," "x mg," "harm attempt," "mg seroquel," "inflicted stab," "suicide attempt," "took x," "polypharm od," "inflicted lacs," "unknown quantity."*

Unigrams + Bigrams:.

> *"self harm," "diazepam," "seroquel," "suicide," "suicidal," "superficial lacs," "od x," "polypharmacy od," "intent," "intentional," "x mg," "attempt," "harm," "superficial," "mg," "tablets," "suicidal intent," "self," "polypharmacy," "od."*

Interestingly, combining bigrams and unigrams did not result in any change in performance (Table 1). Using bigrams alone dramatically improved model's precision but came at a cost of lower recall. This surprising observation prompted us to investigate the effect of bigram detection as described in Supplementary materials.

Table 1. Performance of keyword search using unigrams, bigrams, and a combination of both

	Precision	Recall	F_1 score
Unigrams	0.043 (±0.02)	0.927 (±0.04)	0.081 (±0.03)
Bigrams	0.429 (±0.06)	0.635 (±0.04)	0.512 (±0.05)
Unigrams + bigrams	0.047 (±0.00)	0.895 (±0.04)	0.089 (±0.00)

Note: Results are reported as mean (±95% confidence intervals) obtained from 10-fold cross-validation on the training set

Machine learning classifiers

Machine learning (ML) models differ in their complexity and vary in their ability to capture relationships between features and the outcome variable. We evaluated a suite of ML models including Naive Bayes, Logistic Regression, k-Nearest Neighbors, Random Forest, and Gradient Boosting (GB). We investigated the effect of three key preprocessing steps including bigram replacement, clinical concept recognition, and negation detection (see Supplementary materials for details).

When using BOW, commonly co-occurring words can be taken into account either by performing bigram replacement or by computing the TF-IDF matrix for both unigrams and bigrams. In our experiments, both techniques resulted in an impaired model performance (Supplementary Table S3). We also evaluated concept recognition and negation detection applied either separately or in combination. However, none of these approaches resulted in any significant changes in the model performance (Supplementary Tables S4–S6). As such, below we report the results of training selected ML models on the training set without additional preprocessing steps using unigrams only (Table 2). Notably, the overall performance of the GB algorithm was consistently better than other models as evidenced by a higher PR-AUC score.

Deep learning

To leverage the sequential nature of text we also implemented a long short-term memory (LSTM) network which has the property of remembering previously seen inputs and thus can take into account the order of the words in a sentence.

We compared various configurations of LSTM (Table 2, Supplementary Table S7), however, the performance did not exceed that of the GB model and the training was much more time-consuming. As such, we selected the GB algorithm for further development of the SH classification system.

Model calibration

One of the ways to gauge the reliability of an ML model is to assess whether the distribution of forecasted probabilities matches the expected distribution of observed probabilities. A diagnostic plot presented in Figure 2 illustrates the distributions of probabilities predicted by the calibrated and uncalibrated GB models in comparison to a perfectly calibrated model that would have points strictly along the main diagonal. While the uncalibrated GB model already performed relatively well, its calibration further improved the performance achieving a slightly higher score (PR AUC = 0.839 compared to PR AUC = 0.832).

Evaluation of the classifier on the test set

After tuning hyper-parameters of the calibrated model on the training set, predicted probabilities were converted to class labels by

Table 2. Performance of several ML classifiers on the task of detecting instances of self-harm in the ED triage texts corpus

	PR AUC	Precision	Recall	F_1 score
Naive Bayes	0.666 (±0.03)	0.550 (±0.03)	0.707 (±0.03)	0.618 (±0.02)
Logistic regression	0.799 (±0.03)	0.296 (±0.01)	0.952 (±0.01)	0.452 (±0.01)
k-nearest neighbor	0.463 (±0.05)	0.892 (±0.04)	0.276 (±0.04)	0.421 (±0.05)
Random forest	0.799 (±0.03)	0.843 (±0.03)	0.601 (±0.05)	0.702 (±0.04)
Gradient boosting	0.832 (±0.03)	0.855 (±0.04)	0.691 (±0.05)	0.764 (±0.03)
LSTM*	0.801 (±0.04)	0.560 (±0.08)	0.874 (±0.02)	0.682 (±0.05)

Note: Results are reported as mean (±95% confidence intervals) obtained from 10-fold cross-validation on the training set. For LSTM, we used 3-fold cross-validation.

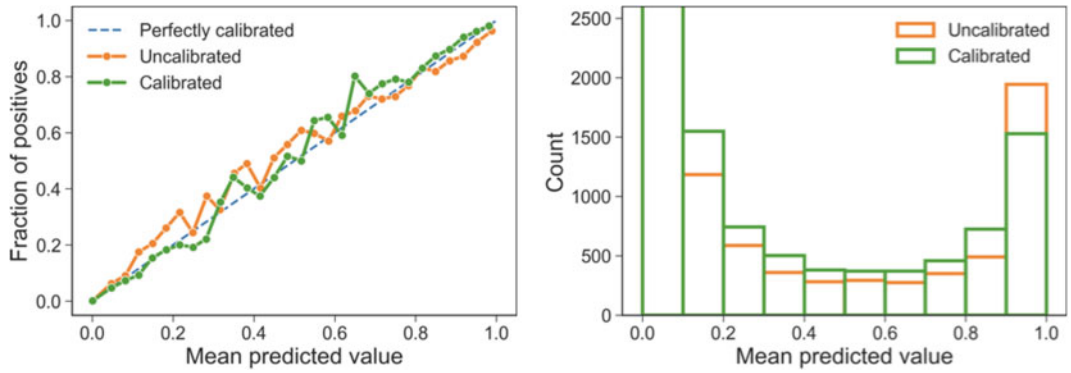

Figure 2. Diagnostic plots illustrating the difference in predictions made by the calibrated and uncalibrated models.

selecting a threshold optimizing the F_1 score. The highest $F_1 = 0.788$ was achieved by setting the threshold to 0.32, meaning each triage note with a probability higher than that was predicted as positive for SH.

The final model was based on retraining the preferred GB method with the complete training set, selecting 970 features. We evaluated this model on the previously unseen test set (20% of the development set held-out for testing). Figure 3 shows the ROC and Precision-Recall curves of this model on the test set demonstrating that the classification system is highly skilled at predicting the negative class and accurate at detecting SH. The model correctly identified 861 positive cases out of 1076 achieving macro Precision = 0.899, Recall = 0.899, and F1 score = 0.899 (Figure 3, Table 3).

Prospective model validation on data from 2018 and evaluation for confusion with suicidal ideation cases

Prospective validation of the developed model was performed by evaluating the classifier on the set of ED presentations from 2018 previously excluded from the development of the model (Figure 1). The number of controls present in the hold-out set was comparable to previous years whereas the number of SH cases appeared to gradually increase (Figure 4). When evaluated on this data, our final model showed similar results with macro Precision = 0.878, Recall = 0.888, and F1 score = 0.883 (see Table 3). Figure 5 shows examples of false-negative triage notes illustrating how each word contributes towards the model's predictions. The triage note presented in Figure 5B, particularly, was predicted as SH with a probability just below the threshold and as such was classified as Control. This example highlights the importance of predicting probabilities rather than crisp classes and paying closer attention to cases when model predictions are uncertain.

We further sought to evaluate whether there was any confusion between SH and SI cases affecting the performance of the classifier. Using the annotations provided for data from 2018, we found that out of 301 false-positive triage notes, 38.5% ($N = 116$) were evident of SI accounting for 7.7% of all SI cases. Examples of SI-positive notes predicted as SH are given in Figure 6A,B. Predictably, some confusion between these categories was caused due to the use of similar vocabulary, such as the word "suicidal" and words related to taking medication. Figure 6C illustrates an example of a false-positive triage note possibly due to the missed negation of the word "intentional."

DISCUSSION

Detection of self-harm (SH) cases presenting to an emergency department (ED) based solely on ICD-10 codes is unreliable due to the high false-negative rate.[14,29] The World Health Organization recommends the use of International Classification of Diseases (ICD) codes to achieve consistency and uniformity in the identification of SH cases.[30] However, administrative data sets, which often form the backbone of case ascertainment protocols within these systems, vary in completeness in ICD coding, particularly for external cause injury codes.[31] As external cause codes are essential for identifying SH cases, systems that rely exclusively on ICD coding are likely to significantly underestimate the true number of SH cases.[11,31] In contrast, the use of free-text triage data has been found to improve the automated detection of SH cases in real-world applications.[32] The strong performance of our model on a large, real-world data set demonstrates the viability of using ED texts to support surveillance of SH, without the delays introduced by reliance on retrospective, manual coding processes.

It is worth noting that any preselection of ED presentation using Primary Diagnostic ICD-10 Codes can also impose the risk of missing positive cases resulting in a model that is less likely to generalize

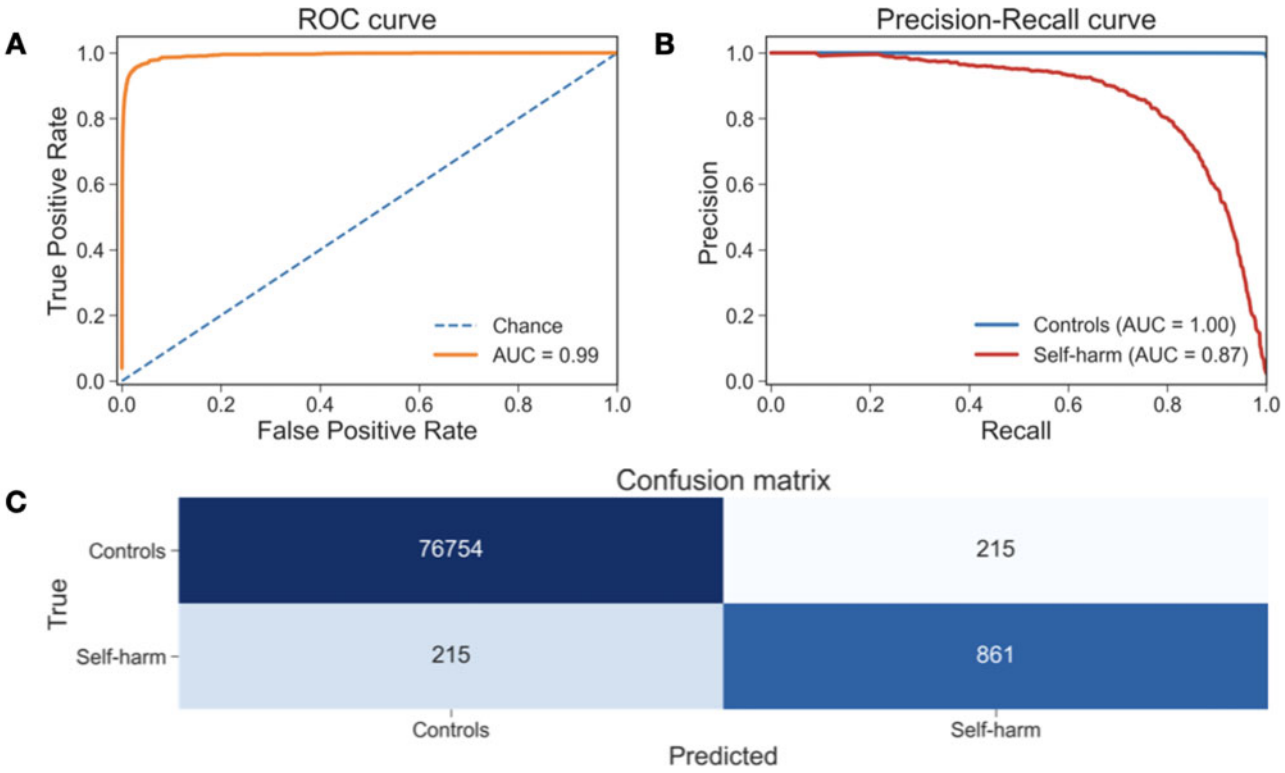

Figure 3. (A, B) ROC and Precision-Recall curves of the final GB classifier when evaluated on the unseen test set. (C) Confusion matrix shows the number of correctly predicted and misclassified cases.

Table 3. Final model predictions on the test and hold-out sets

	Precision	Recall	F_1 score	TN	FP	FN	TP
Test set (2012–2017) $N = 78\,045$	0.899	0.899	0.899	76 754	215	215	861
Hold-out set (2018) $N = 76\,383$	0.878	0.888	0.883	74 864	301	269	949

well.[33] In our study, all ED presentations were reviewed by two annotators with expertise in suicide prevention achieving interannotator agreement of 0.91 for both SH and suicidal ideation (SI) presentations. Our results show that machine learning methods, specifically Gradient Boosting (GB), significantly outperform conventionally employed approaches involving keyword search. Moreover, the list of identified keywords differed considerably from the previously reported application to clinical notes,[20] likely due to the specifics of the ED nursing notes. At the same time, the GB algorithm resulted in equivalent performance to the evaluated neural models characterized by a markedly longer training time.

Needless to say, the annotation of half a million ED triage notes involved a significant amount of effort. Manual coding alone required around 1000 person hours and the total amount of advisory and auditing time added up to 9000 person hours. Recognizing that replicating this component of our work may not be feasible, we are providing the final fitted GB model in Supplementary materials for use in other application contexts (the model and the code are also available on GitHub (https://github.com/vlada-rozova/self-harm-jamia, last accessed November 29, 2021)).

It is worth emphasizing the high imbalance of the RMH data with a mere 1.4% of all presentations being identified as related to SH, making this a very challenging task for any classifier. Approaches to tackling problems with imbalanced data sets include undersampling the majority class (in our case, ED presentations unrelated to SH), which might lead to significant loss of information, and oversampling the minority class, either by replicating or synthesizing positive examples. However, techniques for generating synthetic instances such as SMOTE[34] tend to work poorly with high-dimensional textual data.[35] While more detailed investigation of this issue is warranted, in this study, we employ cost-sensitive algorithms to take into account the prior class distribution. Additionally, it is important to evaluate performance using metrics agnostic to class imbalance. Commonly reported ROC AUC metric is sensitive to such differences in numbers hence in this study we calculate PR AUC to compare and select the best performing model.

Prospective validation of the model using data from 2018 aimed to reproduce a realistic scenario in which a model is developed using data up to a given point in time and then applied to newly collected data. At the same time, the test set was generated by randomly sampling 20% of encounters from the development set spanning years 2012–2017. The fact that there is only a small drop in performance between the test set and the hold-out set (Table 3) indicates that the model generalizes well on prospectively collected data.

Additionally, ED presentations from 2018 annotated for SI allowed for the evaluation of confusion between SI and SH cases. When applied to preidentified SI cases, our model misclassified less than 8% of these as SH. Given the close relationship between SH and SI, a much higher rate of ambiguity might have been expected. Accurately distinguishing between these types of cases

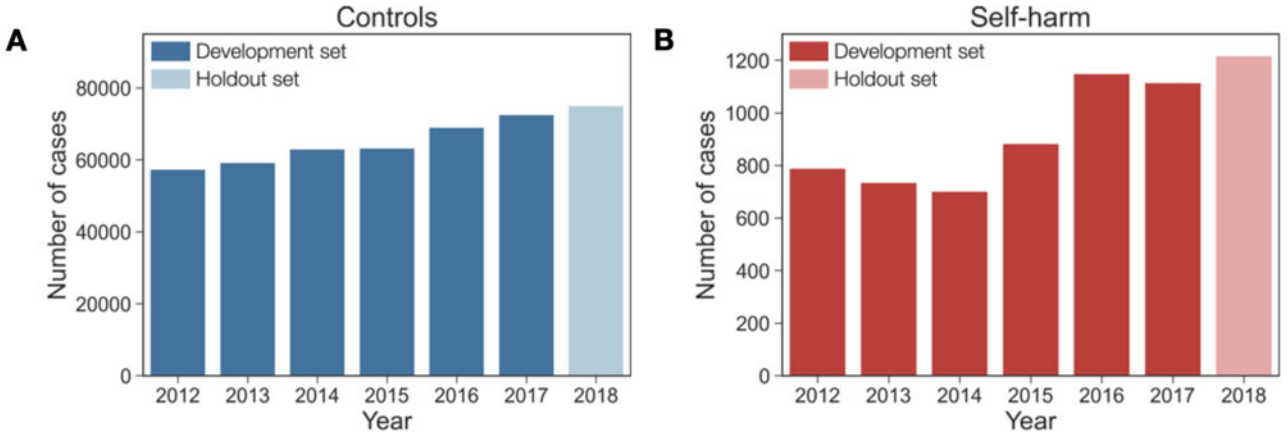

Figure 4. Changes in the numbers of recorded cases negative and positive for SH. Lighter colors in both panels correspond to the hold-out set used for prospective model validation.

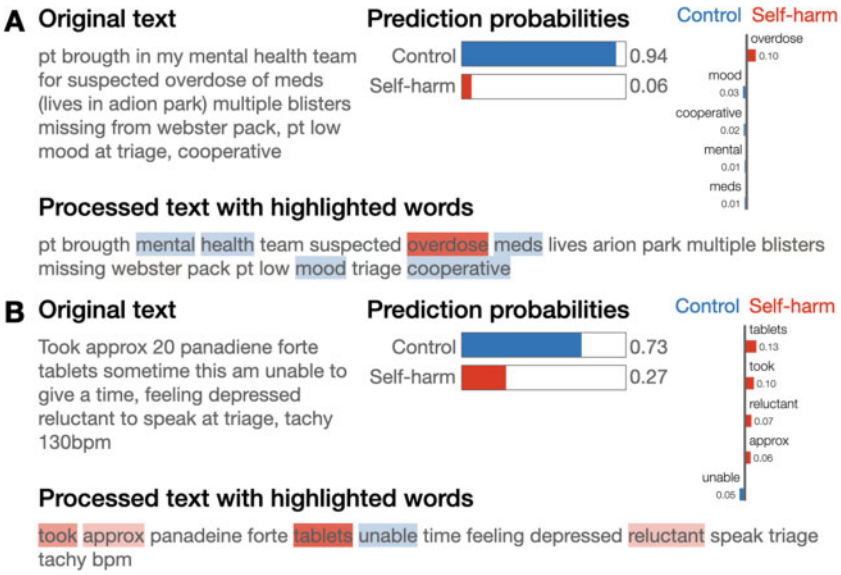

Figure 5. (A, B) Illustration of triage notes annotated as SH and misclassified as Controls (false negatives). The bars on the left show the predicted probability of each class. Horizontal bar plot provides the weights of five most important features. On the bottom, these words are highlighted in the text.

will allow for the identification of different pathways from SI to SH and for the empirical testing of different models within the ideation to enaction framework. Furthermore, nearly 40% of false-positive triage notes were positive for SI. Taken together, these findings suggest that a multiclass classification model could reduce the probability of false alarm for SH while also enabling the detection of SI cases. In future work, we plan to explore the development of a model that explicitly aims to distinguish between SH and SI cases more carefully.

We further have only evaluated the model in the context of a single hospital, which leaves open the question of its relevance to other ED data sets with different data characteristics, including variations in clinical language, and potentially distinct distributions of self-harm or suicidal ideation. Through the collaborations in the context of the state-wide self-harm monitoring system under development in our Australian state of Victoria,[36] we expect to be able to explore the generalization of the model to other hospital contexts.

CONCLUSION

We have developed an automatic self-harm classification system for ED presentations, a novel application of clinical natural language processing (NLP), and the first automated self-harm classification system based directly on ED triage notes. This system was built leveraging a large manually annotated data set of ED nursing triage notes and incorporates a number of NLP steps that are effective for normalization and representation of these relatively short and noisy clinical texts. False positives of the model to a large extent are attributable to confusion between self-harm and suicidal ideation, which we aim to address in future work via a multiclass classification approach. Our model achieved high sensitivity and positive predictive value on this naturally occurring and highly imbalanced data set and is therefore viable both as a surveillance tool and for clinically actionable alerts in ED. We believe that this system will allow for the timely identification of changes in the patterns of suicide in Victoria and will have the capacity to act as barometers of the success of national suicide prevention strategies and to inform

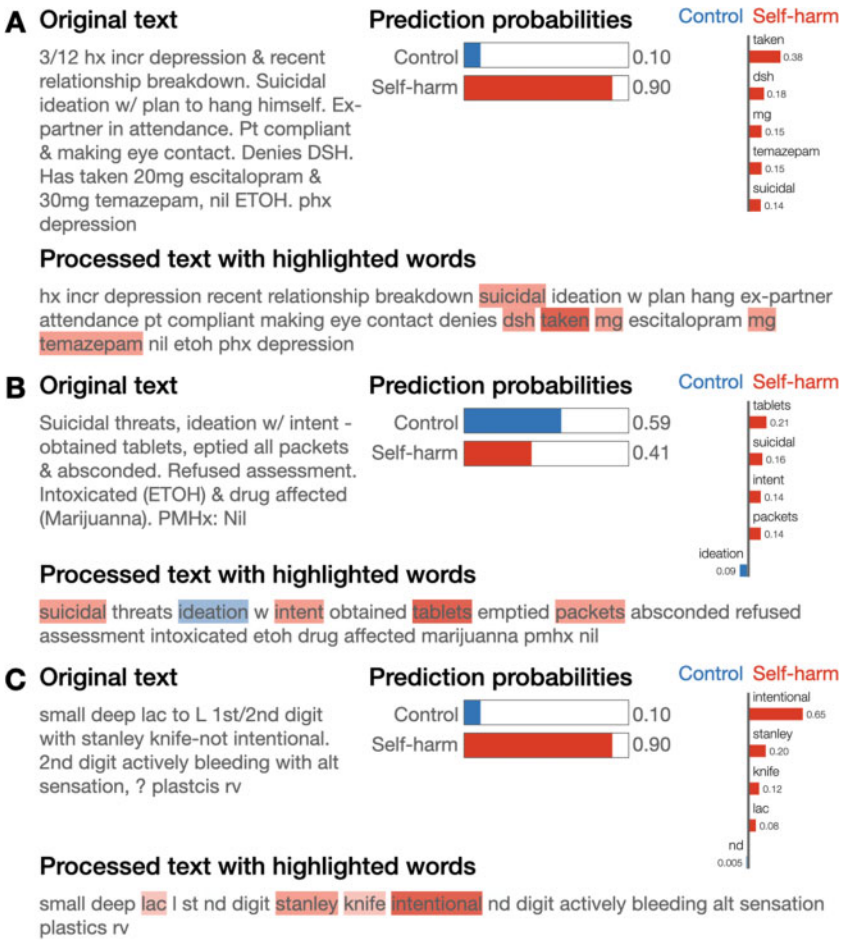

Figure 6. Illustration of triage notes misclassified as SH (false positives). (A, B) 40% of triage notes misclassified as SH were in fact annotated as SI. (C) An example of a triage note negative for both SI and SH but misclassified as SH. The bars on the left show the predicted probability of each class. Horizontal bar plot provides the weights of 5 most important features. On the bottom, these words are highlighted in the text.

real-time change and intervention at both the population and individual levels.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Journal of the American Medical Informatics Association online.

FUNDING

The National Health and Medical Research Council (NHMRC) (1142348 to JR and 1177787 to KW). KV and VR acknowledge support from NHMRC grant 1134919.

AUTHOR CONTRIBUTIONS

VR and KV developed the NLP approach for the modeling and interpreted the results. VR implemented and evaluated the NLP methods. YL contributed to prototyping NLP methods. JR and KW contributed to the conceptualization of the study. KW labeled the data set. VR drafted the manuscript, and VR, KV, KW, and JR revised and finalized the writing. All authors approved the final manuscript.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study. The data will be shared on reasonable request to the corresponding author.

ACKNOWLEDGMENTS

We acknowledge Dr Jonathan Knott of the University of Melbourne and the Royal Melbourne Hospital for providing access to the ED presentation data that was the foundation of this study.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Bachmann S. Epidemiology of suicide and the psychiatric perspective. *Int J Environ Res Public Health* 2018; 15 (7): 1425.
2. Australian Bureau of S. Causes of Death, Australia, 2019. Canberra, ACT: Australian Bureau of Statistics (ABS); 2019.
3. Australian Institute of Health and Welfare. Suicide prevention activities. Canberra, Australia: Australian Institute of Health and Welfare; 2018.
4. Australian Institute of Health and Welfare. Trends in Hospitalized Injury, Australia 1990-00 to 2014-15. Canberra, ACT: AIHW; 2018.

5. Hawton K, Zahl D, Weatherall R. Suicide following deliberate self-harm: Long-term follow-up of patients who present to a general hospital. *Br J Psychiatry* 2003; 182: 537–24.
6. Zahl DL, Hawton K. Repetition of deliberate self-harm and subsequent suicide risk: Long-term follow-up study of 11 583 patients. *Br J Psychiatry* 2004; 185 (1): 70–5.
7. Hawton K, Bergen H, Casey D, *et al.* Self-harm in England: a tale of three cities. Multicentre study of self-harm. *Soc Psychiatry Psychiatr Epidemiol* 2007; 42 (7): 513–21.
8. Perry IJ, Corcoran P, Fitzgerald AP, Keeley HS, Reulbach U, Arensman E. The incidence and repetition of hospital-treated deliberate self harm: findings from the world's first National Registry. *PLoS One* 2012; 7 (2): e31663.
9. Hiles S, Bergen H, Hawton K, Lewin T, Whyte I, Carter G. General hospital-treated self-poisoning in England and Australia: comparison of presentation rates, clinical characteristics and aftercare based on sentinel unit data. *J Psychosom Res* 2015; 78 (4): 356–62.
10. Victorian emergency minimum dataset (VEMD). State of Victoria, Australia: Department of Health; June 2021. https://www.health.vic.gov.au/data-reporting/victorian-emergency-minimum-dataset-vemd
11. Witt K, Robinson J. Sentinel surveillance for self-harm: existing challenges and opportunities for the future. *Crisis* 2019; 40 (1): 1–6.
12. Hawton K, Bergen H, Casey D, *et al.* Self-harm in England: a tale of three cities. *Soc Psychiatry Psychiatric Epidemiol* 2007; 42 (7): 513–21.
13. Müller A, Claes L, Smits D, Brähler E, de Zwaan M. Prevalence and correlates of self-harm in the German general population. *PLoS One* 2016; 11 (6): e0157928.
14. Walkup JT, Townsend L, Crystal S, Olfson M. A systematic review of validated methods for identifying suicide or suicidal ideation using administrative or claims data. *Pharmacoepidemiol Drug Saf* 2012; 21 Suppl 1: 174–82.
15. Choi SB, Lee W, Yoon J-H, Won J-U, Kim DW. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *J Affect Disord* 2018; 231: 8–14.
16. Velupillai S, Suominen H, Liakata M, *et al.* Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. *J Biomed Inform* 2018; 88: 11–9.
17. Graham S, Depp C, Lee EE, *et al.* Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep* 2019; 21 (11): 116.
18. Carson NJ, Mullin B, Sanchez MJ, *et al.* Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PLoS One* 2019; 14 (2): e0211116.
19. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci Rep* 2018; 8 (1): 7426.
20. Obeid JS, Dahne J, Christensen S, *et al.* Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. *JMIR Med Inform* 2020; 8 (7): e17784.
21. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017; 12 (4): e0174708.
22. Gligorijevic D, Stojanovic J, Satz W, Stojkovic I, Schreyer K, Obradovic Z. *Deep Attention Model for Triage of Emergency Department Patients*. ArXiv abs/1804.03240. 2018; 9.
23. Neumann M, King D, Beltagy I, Ammar W. *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing*. Florence, Italy: Association for Computational Linguistics; 2019: 319–327.
24. Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 160035.
25. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records. Artif Intell Med 2021; 118: 102086. doi:10.1016/j.artmed.2021.102086.
26. pyspellchecker 0.6.2 [Internet]. Python Software Foundation; [cited 2021 Nov 29]. https://pypi.org/project/pyspellchecker/
27. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY: Association for Computing Machinery; 2016: 1135–44. (KDD '16). https://doi.org/10.1145/2939672.2939778
28. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594.
29. Anderson HD, Pace WD, Brandt E, *et al.* Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med* 2015; 28 (1): 65–71.
30. World Health Organisation. Practice manual for establishing and maintaining surveillance systems for suicide attempts and self-harm. Geneva, Switzerland: World Health Organization; 2016.
31. Hedegaard H, Schoenbaum M, Claassen C, Crosby A, Holland K, Proescholdbell S. Issues in developing a surveillance case definition for nonfatal suicide attempt and intentional self-harm using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coded data. *Natl Health Stat Report* 2018; (108): 1–19.
32. Sperandei S, Page A, Spittal MJ, Witt K, Robinson J, Pirkis J. Using the 'presenting problem' field in emergency department data improves the enumeration of intentional self-harm in NSW hospital settings. *Aust N Z J Psychiatry* 2020; 55 (10): 1019–20.
33. Stapelberg NJC, Randall M, Sveticic J, Fugelli P, Dave H, Turner K. Data mining of hospital suicidal and self-harm presentation records using a tailored evolutionary algorithm. *Mach Learn Appl* 2021; 3: 100012.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–57.
35. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013; 14 (1): 106.
36. Robinson J, Witt K, Lamblin M, *et al.* Development of a self-harm monitoring system for Victoria. *Int J Environ Res Public Health* 2020; 17 (24): 9385.