Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2023, Vol. 67(1) 1063–1069
Copyright © 2023 Human Factors and Ergonomics Society
DOI: 10.1177/21695067231193645
journals.sagepub.com/home/pro
Sage

Improving Patient Safety Event Report Classification with Machine Learning and Contextual Text Representation

Hongbo Chen[1], Eldan Cohen[1], Dulaney Wilson[2], and Myrtede Alfted[1]

Abstract

Adverse events caused by medical errors pose a significant threat to patient safety, with estimates of 251,454 deaths and a cost of $17.1 billion to the healthcare system annually in the United States. Patient safety event (PSE) reports play a vital role in identifying measures to prevent adverse events, but their utility is dependent on the accurate classification of PSE reports. Recent studies have used static natural language processing (NLP) and machine learning (ML) techniques to automate PSE report classification. However, the use of static NLP has limitations in differentiating the meaning of words in disparate contexts, which can lead to inferior classification results. Thus, this study proposes to utilize contextual text representation produced from neural NLP methods to improve the accuracy of PSE report classification. The results suggest that the contextual text representation can further improve the performance of PSE classifiers. The best-performing classifier, a support vector machine trained with contextual text representation (Roberta-base) reaches an accuracy of 0.75 and a ROCAUC score of 0.94, surpassing all ML classifiers trained with static text representations. Furthermore, the confusion matrix of the best classifier exposes latent deficiencies in the PSE reports' classification taxonomy, such as the multi-class nature of PSE and conceptually related event types. The study's findings can save time for PSE reclassification, enhance the learning capabilities of the reporting system, ultimately improve patient safety

Keywords

Machine Learning, Text Classification, Patient Safety Event

Introduction

Adverse events due to medical errors remain a major threat to patient safety globally (Makary & Daniel, 2016). Adverse events are unintended injuries or complications caused by delivery of care, rather than by the patient's underlying disease. Adverse events can prolong the patient's hospitalization, result in escalation of care or additional treatments, or produce injury or a disability at the time of discharge (Sari et al., 2007). Approximately 4 to 17 % of patients experience adverse events across the world (Rafter et al., 2015). Experts estimate that as many as 251,454 patients die annually due to adverse events in the United States (Makary & Daniel, 2016). The annual cost of adverse events in the United States has been estimated at $17.1 billion (Van Den Bos et al., 2011).

Patient safety event (PSE) reports play a vital role in allowing healthcare organizations to learn from adverse events and develop measures to improve patient safety (Puthumana et al., 2021). A PSE is an event or circumstance which could have resulted or did result in harm to a patient. PSE report consists of structured data (i.e., event types, patient harm level, location of the event) and unstructured data (i.e., free text section for describing the event, patient outcome, and so on) (Fong et al., 2021). PSEs can be reported by any hospital staff member in the hospital's PSE reporting system. The reporter describes the event and completes fields regarding the date, time, location, and potential causes. Once the report is submitted, the event may be reviewed by the hospital's patient safety and risk teams as well as relevant managers and quality supervisors. PSE reports often drive patient safety and quality improvement efforts at a hospital (Puthumana et al., 2021).

[1]Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada
[2]Public Health Sciences, Medical University of South Carolina, Charleston, United States

Corresponding Author:
Hongbo Chen, Mechanical and Industrial Engineering, University of Toronto, 5 king street, Toronto, M5S 1A1, Canada.
Email: hongbo.chen@mail.utoronto.ca

The use of PSE reports to examine specific patient safety issues is highly dependent on classifying PSEs into their correct event types (Wang et al., 2017). An event type is a descriptive term for a class of events with a common nature. PSE reporting systems can have upwards of 20 event types (Evans et al., 2020). Healthcare personnel often struggle with classifying PSEs (Mahajan, 2010). The reporter has to make a subjective decision about the most appropriate event type based on their interpretation of the event (i.e., did the reporter witness the event) and level of understanding of the PSE reporting system's taxonomy (i.e., experience with classification taxonomy). The consistency of classification results tends to vary depending on the event reporter's profession (i.e., nurse, physician, technician) (Lee et al., 2020). As a result, many PSE reports are misclassified. Misclassification limits the PSE reporting system's learning functionality, requires reclassification of PSEs, and potentially confounds the database during pattern searches for developing solutions (Puthumana et al., 2021). Additionally, PSE reports were frequently classified as "miscellaneous", which requires a substantial amount of effort to reclassify into appropriate event types (Fong et al., 2021). In order to secure the most value from PSE reports, hospitals need to ensure the events reported are classified accurately.

To more efficiently and accurately classify PSE reports, various recent studies have utilized static natural language processing (NLP) and machine learning (ML) techniques to automate PSE report classification (Fong et al., 2021; Evans et al., 2020; Wang et al., 2022). The performance of these classifiers is promising and has proven its efficacy in classifying PSE reports. However, previous studies used the text representation produced from static NLP algorithms, which does not differentiate the meaning of the word in disparate contexts (Fong et al., 2021, Evans et al., 2020). This can lead to inferior classification results. Neural NLP algorithms have improved the representation of text documents by using contextual text representation generated from deep learning (DL) language models. These models are usually trained on large amounts of unstructured text in a self-supervised fashion (Liu et al., 2019). The neural NLP algorithms assign each word's representation based on its context, thus enabling capturing of more accurate meanings of words across varied contexts (Liu et al., 2020).

Therefore, the primary objective of this study is to examine the efficacy of contextual text representation coupled with ML classifiers in classifying PSE reports. First, various ML classifiers, including softmax regression, support vector machine, random forests, K-nearest neighbors, extreme gradient boosting, and light gradient boosting, were trained with different static text representations. These ML classifiers were then trained with various contextual text representations. The performance of ML classifiers trained on both static and contextual text representation was compared based on various multi-class classification metrics. Furthermore, we also analyzed the confusion matrix of the top-performing classifier to understand its performance and identify areas of improvement. This work can reveal meaningful insight for improving the reliability of PSE classification in the event reporting system.

Methods

Data Collection

The dataset used to train classifiers was obtained from a large academic hospital in the Southeastern United States. A total of 861 PSE reports from January 1st, 2019, to December 31st, 2020, were extracted from the event reporting system for the labor and delivery (L&D) and mother-baby (MB) units as part of a larger study examining adverse events in maternal care. The study was approved by the hospital's institutional review board (Pro00105892).

Following data extraction, all PSE reports were anonymized for privacy regulation. There were 25 event types in the PSE reporting system including *complications of surgery, falls, medication-related, environmental issues*, and so on. Only the free-text section was used for classifying PSE reports. The PSE reports from seven frequent event types were selected for training classifiers. The decision was made to avoid sampling bias. The included PSE reports' event types and associated frequencies were *care coordination/communication* (186), *laboratory test* (122), *medication-related* (89), *omission/errors in diagnosis, monitoring* (67), *maternal* (58), *equipment/devices* (56), and *supplies* (49). The selected PSE reports accounted for 73% of the reported events.

Data Preparation

The free text section of PSE reports was preprocessed before being used as input for training classifiers. The preprocessing procedures include data cleaning, feature extraction, data splitting, and data augmentation.

Data cleaning. Two types of features, including static text representation and contextual text representation, were extracted from the free text section of PSE reports using static NLP algorithm and neural NLP algorithm, respectively. The static NLP algorithm requires the text to be normalized so that the algorithm can recognize the same word in a different format (i.e., plural vs. singular).

The following text normalization procedures were completed for obtaining static text representation from the static NLP algorithm: changing the case of the word to lowercase, removing non-alphabetical characters, and names, and stemming. Stemming reduces words into their base form, so that algorithm can recognize the same word in different formats. The normalized text allowed the static NLP algorithm to produce a consistent representation of the same word in a different format. The contextual text representation does not require text normalization because the neural NLP algorithm was trained on the raw text that has not been normalized. Neural NLP algorithms are more robust to anomalies such as spelling mistakes, word tenses, and plurality (Devlin et al., 2019; Liu et al., 2019).

Feature extraction. A total of three static text representations were extracted. The normalized text was converted into the bag of words (BOW) format, term frequency-inverse document frequency (TF-IDF) format, and global vectors (Glove) format. BOW is a vector format representation of a text document where the total occurrence of each word in the document is used as a feature for training an ML classifier. TF-IDF is a vector format representation of a text document that reflects how relevant a word is to a text document in the entire dataset. The static text representation produced from BOW and TF-IDF in this study all used n-grams ranging from 1 to 3. The n-gram refers to how we tokenize text into different parts. which is necessary to represent text in a way that the computer can read. Glove algorithms represent each word with a vector that captures the semantic relationship between words. The representation of the entire text document is then obtained by aggregating each word's vector.

A total of six contextual text representations were extracted by passing the event description into neural NLP algorithms, including Bert-cased, Bert-uncased, xlm-Roberta-base, Roberta-large, Roberta-base, and PubMed-Bert. Each word is represented with a vector that captures the contextual and positional information of itself within the text document. The text document's representation is obtained by aggregating words' vectors and dividing by the number of words in that text document. The difference between the six contextual text representations in this study is coming from the domain of text on which the neural NLP algorithm is trained.

Data splitting. For both static text representations and contextual text representations of PSE reports, data was splitted with stratified sampling with respect to the frequency of event types; 80% of the data was used for training and 20% was used for testing. The classifiers were trained with the training data and evaluated with the testing data, which is a standard approach in the ML field. This was done to avoid overfitting. Additionally, splitting in a stratified fashion allows the data in the training set and testing set to preserve the original class's distribution and ensures both the training and testing dataset remain representative of real-world scenarios. We did not create a validation set, instead, the five-fold cross-validation was implemented during hyperparameter tuning to provide classifiers with more access to data during training.

Data augmentation. The data used in this study was imbalanced, and this has the potential to compromise the performance of classifiers. For instance, a previous study noted that imbalanced data can cause K-Nearest Neighbor classifiers to be biased toward the majority class (Kumar et al., 2021). In this study, we used the synthetic minority oversampling technique (SMOTE) to address the problem of imbalanced data. SMOTE is an oversampling technique where the synthetic data are created for the minority class to achieve a balanced distribution of classes, thus improving classifiers' sensitivity to the minority class (Chawla et al., 2002). Only the training data set was augmented with SMOTE, the testing set maintained its original distribution.

Classifier Development

The ML classifiers used to classify PSE reports include logistic regression (LR), support vector machine (SVM), extreme gradient boosting (XGB), light gradient boosting (LGB), random forest (RF), and K-nearest neighbor (KNN). Although SVM is a binary classifier, it can be used to perform multi-class classification with a one vs one strategy, which treats multi-class classification problem as a series of binary classification problems, taking the number of classes (N), and building N*(N-1)/2 binary classifiers for each pair of classes. The final classification is based on the majority vote of all the binary classifiers. The XGB, LGB, and RF are all tree-based ensemble algorithms that are frequently used for text classification problems (Evans et al., 2020). KNN classifiers predict with a majority voting principle, where the data is classified based on its nearest neighbors' classes.

Hyperparameter tuning. For ML classifiers, the hyperparameters were tuned using the 5-fold cross-validation grid search technique. During this process, a range of values for important hyperparameters are evaluated using 5-fold cross-validation, and the final hyperparameter setting is selected based on the cross-validation performance.

Evaluation of classifier performance. We evaluated the performance of the trained classifiers on the testing set based on the following evaluation metrics: accuracy, F1 score, and area under the receiver operating characteristic curve (AUCROC). Accuracy measures the overall percentage of PSE reports that a classifier correctly classifies; F1 score is the harmonic mean of the precision and recall, which gives a whole picture of a classifier's performance on both precision and recall; precision answers how many PSE reports are classified as one specific class belong to that class; recall represents the proportion of PSE reports that are correctly classified as its true class; AUCROC measures classifies' ability to distinguish between classes. Each of these metrics provides a different perspective on classifiers' performance, and together they give a complete picture of how well the classifier is working.

The performance metrics are computed in a macro average way. For instance, the macro F1 score is computed by summing every class's F1 score and dividing by the number

Table 1. The accuracy of different classifiers on the test set.

Accuracy	**LR**	**SVM**	**XGB**	**RF**	**KNN**	**LGB**
BOW	0.55	0.56	0.62	0.64	0.37	0.56
TF-IDF	0.67	0.65	0.64	0.59	0.48	0.64
Glove	0.53	0.63	0.54	0.60	0.57	0.55
Bert-cased	0.63	0.61	0.52	0.52	0.43	0.56
Bert-uncased	0.68	0.70	0.71	0.64	0.54	0.70
xlm-Roberta-base	0.67	**0.75**	0.67	0.66	0.50	0.70
Roberta-base	0.70	0.60	0.60	0.57	0.44	0.56
Roberta-large	0.56	0.66	0.60	0.62	0.52	0.62
PubMed-Bert	0.68	0.70	0.71	0.64	0.54	0.69

Table 2. The F1 score of different classifiers on the test set.

F1 score	**LR**	**SVM**	**XGB**	**RF**	**KNN**	**LGB**
BOW	0.59	0.57	0.63	0.63	0.39	0.57
TF-IDF	0.63	0.62	0.61	0.56	0.40	0.58
Glove	0.55	0.64	0.52	0.58	0.55	0.58
Bert-cased	0.52	0.51	0.46	0.57	0.46	0.50
Bert-uncased	0.62	0.62	0.50	0.63	0.48	0.59
xlm-Roberta-base	0.68	0.69	0.71	0.64	0.54	0.68
Roberta-base	0.68	**0.75**	0.66	0.65	0.48	0.69
Roberta-large	0.70	0.60	0.53	0.51	0.48	0.60
PubMed-Bert	0.59	0.63	0.55	0.53	0.56	0.60

of classes. Macro average is preferred over micro and weighed average because it treats each class with equal importance, and ensures the classifiers' generalizability across all event types.

To examine the efficacy of different text representations on classifiers' performance. The mean and 95% confidence interval of the classifier's performance across different text representations on F1 and AUCROC were also computed. The confidence interval was computed with the mean and standard deviation of a specific classifier's performance across different text representations. Furthermore, we used a confusion matrix to analyze the performance of the top-performing classifier, aiming to identify specific mistakes made by the classifier and the potential causes of these mistakes.

Results

The ML classifiers' performance metrics on static and contextual text representations are reported in Tables 1 (accuracy), 2 (F1 score), and 3 (AUCROC). Overall, the SVM classifier trained with the Roberta-base representation demonstrated exceptional performance, achieving the highest accuracy, F1, and AUCROC among all classifiers with scores of 0.75, 0.75, and 0.94, respectively.

The performance of each ML classifiers' F1 score and AUCROC across static and contextual text representation is shown in Figure 1. The mean of each classifier's performance metrics on various text representations is demonstrated with the bar chart, and the 95% confidence interval around the mean is represented with an error bar. SVM trained with Roberta-base was the most promising ML classifier in terms of F1 score (0.75) and accuracy (0.75), and the KNN classifier trained with BOW has the lowest F1 (0.39) and accuracy (0.37) among all ML classifiers. In terms of the AUCROC score, all classifiers had similar performance (0.83-0.94), however, the KNN classifiers have shown a comparatively lower performance (0.76-0.83).

The performance of each text representation across different classifiers is presented in Figure 2. In terms of F1 score, two contextual text representations, xlm-Roberta-base (highest F1 = 0.71) and Roberta-base (highest F1 = 0.75), showed superior performance when compared to other text representations. The rest of the text representations did not differ much except the Bert-cased text representation (highest F1 = 0.57), which performed poorly. The confusion matrix for the best-performing classifier, SVM trained with Roberta-base, evaluated on the test set has been shown in Figure 3. The diagonal values represent the PSE reports that have been classified as the actual class, whereas the off-diagonal values are the number of PSE reports that have been wrongly classified. While the classifier was able to classify the majority of PSE reports correctly (95 out of 126), two classes of PSE reports including *omission/errors*

Table 3. The AUCROC score of different classifiers on the test set.

AUCROC score	**LR**	**SVM**	**XGB**	**RF**	**KNN**	**LGB**
BOW	0.84	0.81	0.86	0.88	0.76	0.86
TF-IDF	0.91	0.91	0.87	0.89	0.79	0.89
Glove	0.86	0.91	0.86	0.89	0.78	0.86
Bert-cased	0.83	0.84	0.81	0.86	0.74	0.83
Bert-uncased	0.90	0.90	0.84	0.87	0.80	0.86
xlm-Roberta-base	0.91	0.92	0.91	0.89	0.83	0.91
Roberta-base	0.92	**0.94**	0.92	0.91	0.82	0.93
Roberta-large	0.91	0.91	0.89	0.86	0.78	0.87
PubMed-Bert	0.90	0.84	0.88	0.87	0.83	0.89

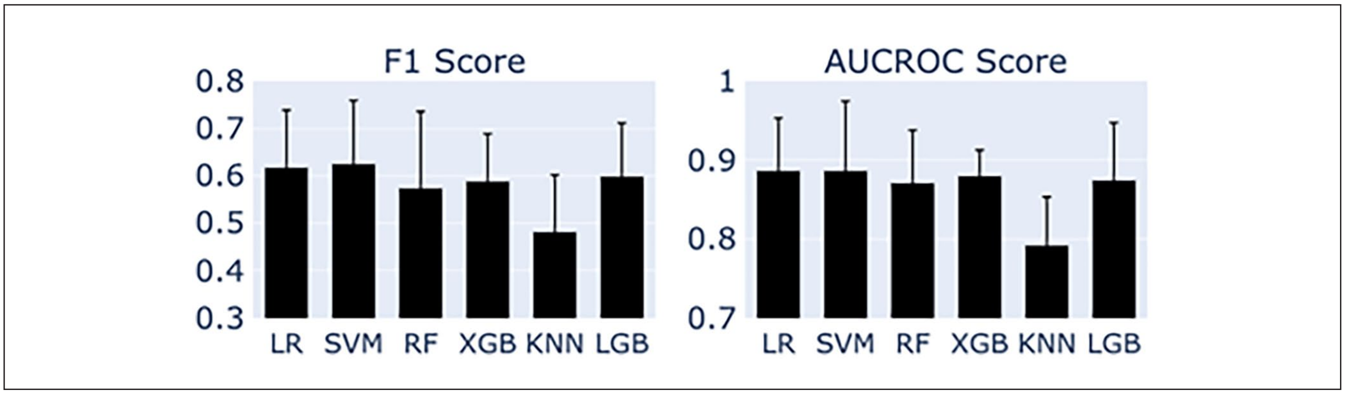

Figure 1. Bar chart of ML classifiers' performance across various text representations.

in diagnosis, monitoring, and *medication-related* events were frequently misclassified.

Discussion

Ensuring PSE reports are correctly classified increases the overall utility of the event reporting system. Incorrect classifications are common and patient safety analysts often have to reclassify incorrectly classified PSE reports and reports classified as miscellaneous (Fong et al., 2021). To improve the efficiency of the process, prior research has sought to automate PSE reports classification using ML classifiers and static text representation (Fong et al., 2021; Evans et al., 2020). This work builds on these methods by utilizing contextual text representation instead of static text representation to enhance the accuracy of PSE report classification.

The best-performing classifier trained with the static text representation (SVM trained with Glove) was able to achieve an accuracy of 0.67, significantly outperforming the baseline accuracy of 0.30 (the accuracy that would have been achieved by classifying every PSE reports as the majority PSE report's event type, 37 out of 125). However, we found the usage of contextual text representation yielded better classification results. The SVM trained with contextual text representation (Roberta-base) was able to achieve an accuracy of 0.75, reflecting an 8% enhancement in accuracy compared to the best-performing classifier trained with the static text representation. The improvement observed in the performance can be attributed to the ability of the contextual text representation to capture the complex and subtle ways in which words interact with each other in different contexts (Liu et al., 2019), thus providing classifiers with a richer and more comprehensive understanding of the text.

Although the classifier was able to differentiate five out of seven event types correctly most of the time, our analysis found two specific events *omission/errors in diagnosis, monitoring*, and *medication-related* events were the most commonly misclassified event types. There are two potential reasons for the misclassification of these PSE reports. First, an individual PSE can be related to multiple event types. For instance, *medication-related* events can originate from insufficient *care coordination/communication* between healthcare personnel, and the inclusion of both causes (*care coordination/communication*) and outcomes (*medication-related error*) in the event report system's taxonomy likely contributes to confusion in selecting the most appropriate event type. Secondly, certain event types are more conceptually related than others. Our confusion matrix showed that *omission/errors in diagnosis, monitoring*, and *care coordination/communication* PSE reports were frequently

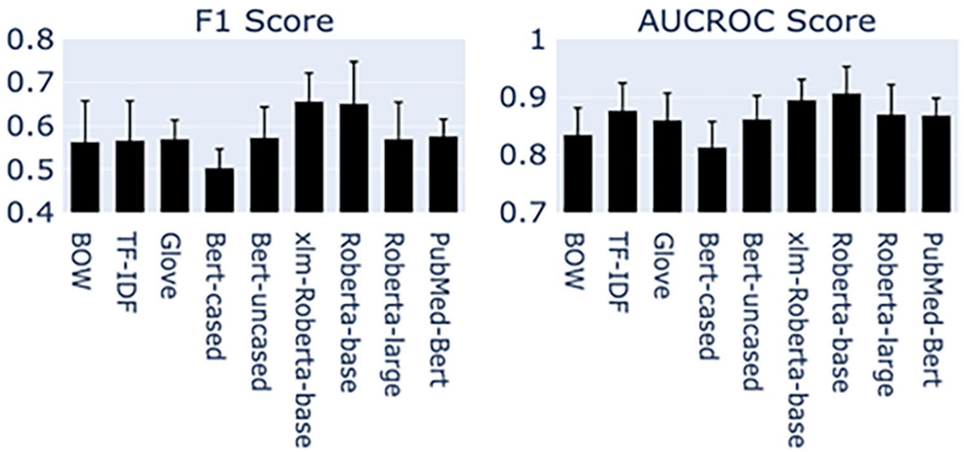

Figure 2. Bar chart of different text representations' performance across various classifiers.

misclassified as one another. Within hospitals, these events likely co-occur. For example, failure to document the removal of a patient's epidural (*omission/errors in diagnosis, monitoring*), prevents the pharmacy from releasing a medication ordered by the physician due to a drug interaction (*care coordination/communication*). However, *laboratory test* PSEs were a more distinct event type compared with other event types, thus the classifier was able to correctly classify the majority of these PSE reports. The observation obtained from the confusion matrix implies that the *omission/errors in diagnosis, monitoring*, and *care coordination/communication* PSE event types' taxonomy definition might be hard to distinguish.

Event reporting systems can have upwards of 20 event types, and healthcare personnel reporting PSEs may not be familiar with event types (Evans et al., 2020). One potential solution could be to modify the taxonomy to make individual event types more conceptually distinct and provide reporters with definitions and examples. Alternatively, the PSE reporting system could be adjusted to allow PSE reports to possess multiple event types. Implementing an ML classifier that prompts top probable event types and allows users to select the most appropriate event type could also facilitate the PSE report classification process.

While our study found promising results, the ML classifiers were trained on PSE reports from a single US hospital. Future research should evaluate the performance of utilizing contextual text representation for training ML classifiers on a more diverse dataset to ensure generalizability. Furthermore, to ensure the reliability of PSE report classification results, explainability techniques should be incorporated to help PSE reporters and patient safety analysts understand the ML classifiers' decisions. As training the ML classifier is just a first step, future research will need to identify and evaluate various strategies for integrating the ML classifier into the workflow of PSE report classifications to support correct classification during reporting and reduce the need for reclassification.

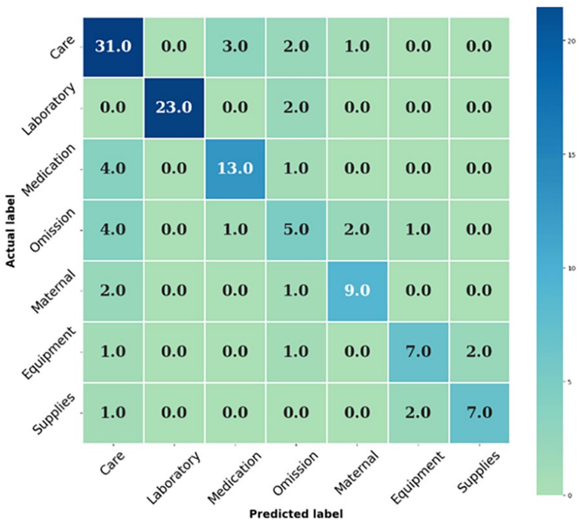

Figure 3. Confusion matrix for testing set evaluation with SVM classifier trained on Roberta-base text representation.

Conclusion

The findings of this study can help advance the development of classification tools for PSE reports. We have demonstrated that the SVM trained with contextual text representation (Roberta-base) provides superior classification results compared with other text representations. Having a PSE reporting system equipped with a built-in feature that can automatically classify the event reports or provide recommendations to reporters can help relieve healthcare personnel's burden on memorizing complicated classification taxonomy and reduce the time spent on reclassifying PSE reports. Additionally, patient safety analysts will spend less time reviewing miscellaneous PSE reports. Based on the findings from the confusion matrix, meaningful insight can be utilized to improve the event type taxonomy. Our next steps include testing the classifiers on a larger PSE report dataset and investigating integration opportunities with an event reporting system. Overall, this work will contribute to establishing a more user-friendly event reporting system and ultimately optimizing organizational learning within health systems.

Acknowledgments

This research was funded by the Agency for Healthcare Research and Quality (AHQR) [Grant no. 1R03HS027680]. The authors declare no conflict of interest.

ORCID iDs

Hongbo Chen https://orcid.org/0009-0005-5823-9406
Eldan Cohen https://orcid.org/0000-0001-5767-6683

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Evans, H. P., Anastasiou, A., Edwards, A., Hibbert, P., Makeham, M., Luz, S., Sheikh, A., Donaldson, L., & Carson-Stevens, A. (2020). Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. *Health Informatics Journal*, *26*(4), 3123–3139. https://doi.org/10.1177/1460458219833102

Fong, A., Behzad, S., Pruitt, Z., & Ratwani, R. M. (2021). A Machine Learning Approach to Reclassifying Miscellaneous Patient Safety Event Reports. *Journal of Patient Safety*, *17*(8), e829–e833. https://doi.org/10.1097/PTS.0000000000000731

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077. https://doi.org/10.1088/1757-899X/1099/1/012077

Lee, K., Yoon, K., Yoon, B., & Shin, E. (2020). Differences in the perception of harm assessment among nurses in the patient safety classification system. *PLoS ONE*, *15*(12), e0243583. https://doi.org/10.1371/journal.pone.0243583

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). *A Survey on Contextual Embeddings* (arXiv:2003.07278). arXiv. https://doi.org/10.48550/arXiv.2003.07278

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv: 1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Makary, M. A., & Daniel, M. (2016). Medical error-the third leading cause of death in the US. *BMJ (Clinical Research Ed.)*, *353*, i2139. https://doi.org/10.1136/bmj.i2139

Puthumana, J. S., Fong, A., Blumenthal, J., & Ratwani, R. M. (2021). Making Patient Safety Event Data Actionable: Understanding Patient Safety Analyst Needs. *Journal of Patient Safety*, *17*(6), e509–e514. https://doi.org/10.1097/PTS.0000000000000400

Sari, A. B.-A., Sheldon, T. A., Cracknell, A., Turnbull, A., Dobson, Y., Grant, C., Gray, W., & Richardson, A. (2007). Extent, nature and consequences of adverse events: Results of a retrospective casenote review in a large NHS hospital. *BMJ Quality & Safety*, *16*(6), 434–439. https://doi.org/10.1136/qshc.2006.021154

Rafter, N., Hickey, A., Condell, S., Conroy, R., O'Connor, P., Vaughan, D., & Williams, D. (2015). 2. *Adverse events in healthcare: Learning from mistakes*. https://doi.org/10.1093/qjmed/hcu145

Van Den Bos, J., Rustagi, K., Gray, T., Halford, M., Ziemkiewicz, E., & Shreve, J. (2011). 3. The $17.1 billion problem: The annual cost of measurable medical errors. *Health Affairs (Project Hope)*, *30*(4), 596–603. https://doi.org/10.1377/hlthaff.2011.0084

Wang, L., Zhang, Y., Chignell, M., Shan, B., Sheehan, K. A., Razak, F., & Verma, A. (2022). Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study. *JMIR Medical Informatics*, *10*(12), e38161. https://doi.org/10.2196/38161