BMJ Health & Care Informatics

A natural language processing approach to categorise contributing factors from patient safety event reports

Azade Tabaie [1], Srijan Sengupta,[2] Zoe M Pruitt [3], Allan Fong [1]

To cite: Tabaie A, Sengupta S, Pruitt ZM, *et al*. A natural language processing approach to categorise contributing factors from patient safety event reports. *BMJ Health Care Inform* 2023;**30**:e100731. doi:10.1136/bmjhci-2022-100731

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100731).

Received 29 December 2022
Accepted 12 May 2023

[1]Center for Biostatistics, Informatics, and Data Science, MedStar Health Research Institute, Washington, District of Columbia, USA
[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA
[3]National Center for Human Factors in Healthcare, MedStar Health Research Institute, Washington, District of Columbia, USA

Correspondence to
Dr Azade Tabaie;
azade.tabaie@medstar.net

ABSTRACT

Objectives The objective of this study was to explore the use of natural language processing (NLP) algorithm to categorise contributing factors from patient safety event (PSE). Contributing factors are elements in the healthcare process (eg, communication failures) that instigate an event or allow an event to occur. Contributing factors can be used to further investigate why safety events occurred.

Methods We used 10 years of self-reported PSE reports from a multihospital healthcare system in the USA. Reports were first selected by event date. We calculated χ^2 values for each ngram in the bag-of-words then selected N ngrams with the highest χ^2 values. Then, PSE reports were filtered to only include the sentences containing the selected ngrams. Such sentences were called information-rich sentences. We compared two feature extraction techniques from free-text data: (1) baseline bag-of-words features and (2) features from information-rich sentences. Three machine learning algorithms were used to categorise five contributing factors representing sociotechnical errors: communication/hand-off failure, technology issue, policy/procedure issue, distractions/interruptions and lapse/slip. We trained 15 binary classifiers (five contributing factors * three machine learning models). The models' performances were evaluated according to the area under the precision-recall curve (AUPRC), precision, recall, and F1-score.

Results Applying the information-rich sentence selection algorithm boosted the contributing factor categorisation performance. Comparing the AUPRCs, the proposed NLP approach improved the categorisation performance of two and achieved comparable results with baseline in categorising three contributing factors.

Conclusions Information-rich sentence selection can be incorporated to extract the sentences in free-text event narratives in which the contributing factor information is embedded.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Contributing factors are important in patient safety event (PSE) report analysis, but the language associated with contributing factors could be subtle and might be embedded in other statements. This makes extracting contributing factors challenging through either manual analysis or machine learning approaches.

WHAT THIS STUDY ADDS

We explored the use of a natural language processing algorithm leveraging the unstructured PSE reports to identify information-rich sentences and demonstrated how this method improved classification performance of contributing factors in PSE reports.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

This approach can be used in near real-time to reduce the burden of manually extracting the factors influencing a patient's safety incident.

INTRODUCTION

Patient safety event (PSE) reporting systems aim to identify safety hazards by encouraging hospital staff to report on errors and potential errors in the hospital system.[1 2] Although PSE reports are limited in that they are often voluntary and only captures a small percentage of the actual prevalence of hazards, these reports have been demonstrated to still be a valuable lens to understand and improve patient safety.[3–6]

PSE reporting systems collect structured and unstructured data. The unstructured data include information about events, such as the contributing factors (CFs) relating to events and patient condition.[7] CFs are important as they represent the factors influencing patient safety incidents (eg, sociotechnical issues, communication failures, technology issues).[8–10] Although identifying and mitigating CFs could improve patient safety, the language associated with CFs could be subtle and difficult to extract as CFs might not always be explicitly described as CFs. It could be interjected between other statements, which makes extracting CFs and using current document-level natural language processing (NLP) and machine learning approaches challenging and often relies on time-intensive manual review. In the example below, while the CF distraction, there is only one sentence about a distraction:

'Registered nurse (RN) was preparing patient for left eye surgery, verified site and

procedure, consent for left eye surgery signed by the patient at bedside. RN was interrupted and inadvertently placed the eye drop for the procedure into the right eye. Patient was notified. Doctor was at bedside'.

The contributions of this work are twofold. First, we explored the use of NLP techniques to categorise CFs from PSE reports. Specifically, we investigated the utility of identifying information-rich ngrams and sentences in categorising CFs. Second, we employed three machine learning algorithms to categorise CFs: logistic regression with elastic net regularisation (elastic net), XGBoost and feed-forward neural network (FFNN).

BACKGROUND

PSE reports

PSE reports contain information regarding adverse events and errors in healthcare.[11] PSE reports contain both structured and unstructured data. For example, the relevant department and level of patient harm are reported as structured data. The event narrative is reported as unstructured, free-text data. While reporting systems encourage reporters to annotate reports with structured, easily searchable data, there are known limitations to reporting systems. They often rely on self-reporting, only captures a small per cent of hazards, can sometimes be bias based on who or what departments are reporting.[12 13] Also, the definition of taxonomies can be confusing.[14]

An example of this is the annotation or coding of CFs. Although reporting systems can give checkbox options to reporters to select associated CFs, they are used infrequently. As a result, relevant information about CFs would only occur in the free-text event narratives.

CFs in PSE

CFs are elements in the healthcare process (eg, sociotechnical issues, communication failures) that instigate an event or allow an event to occur. Human factor models, such as Systems Engineering Initiative for Patient Safety and Human Factors Analysis and Classification System were developed to categorise CFs.[15 16] These CFs can be used to understand changes that need to be made to the system or further investigate why events occurred (eg, interviews, observations).[7]

Challenges with identifying CFs

Although CFs can be selected from a predefined list by reporters, PSE reports are often recorded without a CF indicated in the structured field.[17] Instead, these factors are often described in free-text narratives. Extracting CFs from free text can be challenging because CFs are often interspersed with other text, requiring a time-consuming manual review to extract the information.

Natural language processing

NLP is an algorithmic method for extracting relevant information from free text. In this study, we hypothesise that an NLP approach that uses a sentence selection strategy will successfully identify CFs.

In generating features from free text, the standard options are to either use the term frequency-inverse document frequency (TF-IDF) matrix or word embedding techniques.[18–21] Using bag-of-words and its associated TF-IDF matrix for a categorisation task leads to a high-dimensional feature space that requires a strong computational capacity to train a model. Moreover, the less informative features can add noise to the free-text data and lead to less accurate model performance. On the other hand, while popularly used word-embedding models such as word2vec can help reduce the dimensionality of the problem, utilising word embeddings comes with a loss of interpretability since the original terms are replaced by numeric vectors.[15] Our proposed approach was inspired by previous work in identifying important sentence in free-text categorisation.[22 23] In this study, we hypothesised that an NLP approach, such as a sentence selection algorithm, could be used as a remedy by enabling noise reduction by filtering out the less informative parts of a free-text while preserving interpretability.

METHODS

We explored using an NLP approach to select information-rich sentences relating to CFs information. Then, we used three machine learning algorithms to categorise five sociotechnical CFs. Finally, the effect of the proposed methods on categorisation performance was assessed. Figure 1 demonstrates a summary of the methods used in this study. The Institutional Review Board approved this study.

Data and CF description

The self-reported PSE reports from November 2011 to October 2021 from a multihospital healthcare system in the mid-Atlantic region of the USA were included in this study. In the reporting system, reporters can select over 20 CF options from a list. The CFs are presented to the reporter as a checkbox. Reporters can select none, one or multiple CFs. For this study, we used reports with at least one CF selected by the user to have ground truth for all the included PSE reports. The list of reported CFs and a free-text brief factual description of the event were extracted for each PSE report.

Contributing factors

This study focused on five labels of reported PSE CFs associated with sociotechnical errors: communication/hand-off failure, technology issue, policy/procedure issue, distractions/interruptions and lapse/slip. These five CFs were among our data set's 10 most frequent CFs. Communication/hand-off failure refers to the problems with shift change, patient transfers and information exchange between providers. Technology issues refer to problems with health information technology and medical devices. Policy/procedure issues refer to

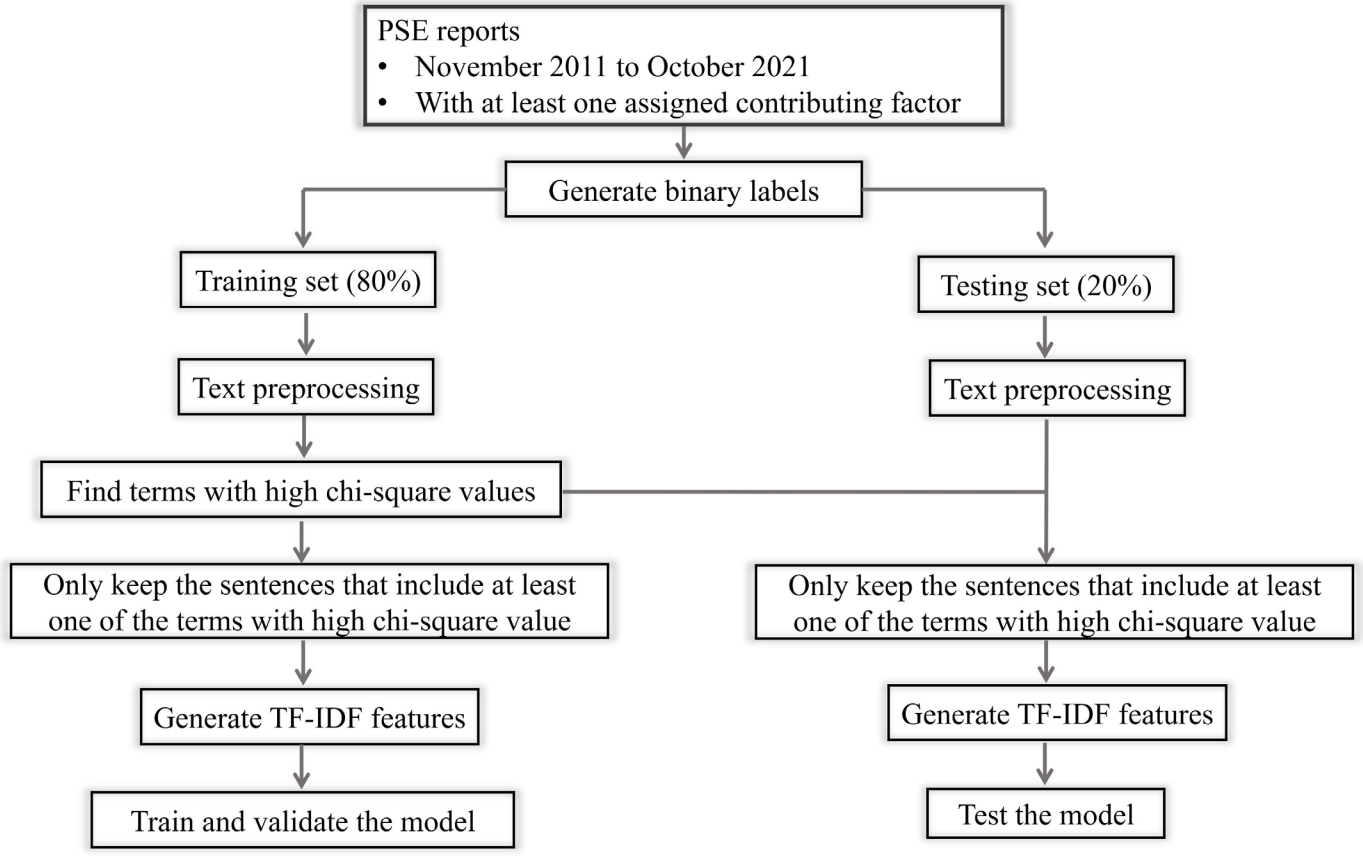

Figure 1 The summary of the methods. PSE, patient safety event; TF-IDF, term frequency-inverse document frequency.

confusing, absent or inappropriate guidelines. Distraction/interruption refers to issues when providers are diverted to a second task before completing the initial task. Lapse/slip refers to issues in human performance, such as accidentally pushing the wrong button.[24]

Text preprocessing

A PSE report can contain multiple CFs. A binary label (ie, one or zero) was assigned to each PSE report before categorising each CF. Text preprocessing is explained in online supplemental appendix A.

Selecting Information-rich terms and sentences

Our proposed NLP approach starts with identifying the information-rich ngrams in each categorisation task. We calculated χ^2 value (for every ngram that was identified in the bag-of-words free-text preprocessing step). χ^2 was calculated using the two-way contingency table of a ngram (t) and a CF (c). In Equation 1, A is the number of times t and c co-occur, B is the number of times t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs and N is the total number of PSE reports in the cohort.

$$X^2 = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

Equation 1. χ^2 calculation to identify information-rich ngrams.

χ^2 measures the degree of association between a specific ngram and the outcome label. This association can indicate a positive or negative relationship between an ngram and a CF; therefore, information-rich ngrams were the ones that achieved a higher χ^2.[25] This approach is motivated by previous work utilising χ^2.[26–28] χ^2 is computationally fast, and the results are easily interpretable.

Finally, selected sentences have at least one of the information-rich ngrams. These sentences will be referred to as 'information-rich sentences'. Concatenating all the information-rich sentences for a PSE report, we produced the free-text for generating the feature matrix.

Machine learning models

Using the bag-of-words from the preprocessed reports, we calculated the TF-IDF matrix associated with the PSE reports. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents and it is calculated by multiplying two metrics: the number of times a word appeared in a document and the inverse document frequency of the word across a set of documents. TF-IDF is a popular method to translate free-text to numerical features in training machine learning models. The data were split into a training set (80%) and a testing set (20%) using stratified sampling. All the preprocessing steps were then applied to the training set, and the same bag-of-words

was incorporated to calculate the TF-IDF matrix of the PSE reports in the testing data.

To assess the effect of the sentence selection method, we used three machine learning strategies (elastic net, XGBoost and FFNN) to categorise the PSE reports and trained separate binary categorisation models for each of the five CFs. Multiple sociotechnical CFs could be assigned to a PSE report; therefore, training a multi-label classification is possible. However, the information-rich sentence selection approach selects different sets of information-rich ngrams for each CF leading to different feature matrices for each classifier; therefore, we trained separate binary categorisation models for each CF. The top N information-rich ngrams were identified through χ^2 calculation for each categorisation task. We set N values as 2, 5, 10, 40, 60 and 100. We compared the performance of these models with their associated bag-of-words, baseline models in which no sentence selection algorithm was applied.

Elastic net

We employed a logistic regression model with elastic net regularisation, which is a weighted combination of least absolute shrinkage and selection operator (LASSO or L1) and ridge (L2) regularisations.[29] Elastic net can remove the effect of the insignificant features by setting their estimated coefficient to zero and lower the effect of the less significant features by pushing their estimated coefficient towards zero while adding more weights to the more important features. Elastic net model is easy to implement and does not require high computation power. Such characteristics make this model an accepted baseline in machine learning-based studies.[30 31] We used elastic net as the benchmark model and compared its results with more complex categorisation methods.

XGBoost

This model is a decision tree-based boosting ensemble machine learning algorithm.[32] In a boosting algorithm, many weak learners are trained to correctly categorise the observations incorrectly classified in the previous training rounds. XGBoost uses a shallow tree as a weak learner and proved to have a decent performance in the case of class-imbalanced data classification.[31 33]

Feed-forward neural network

The feed-forward model is a simple form of a neural network as information is only processed in one direction, and the connection between nodes does not form a cycle.[34] The main benefit of this model is that FFNN accounts for higher order interactions among the input features.[31] We used one hidden layer, a binary cross-entropy loss function and Adam optimiser to train this model.[35]

Instead of data-level solutions (Synthetic Minority Oversampling TEchnique (SMOTE),[36] under-sampling, oversampling), we incorporated algorithm-level solutions (eg, boosting methods, neural networks) as remedy to the class-imabalanced problem in this categorisation problem. To evaluate the performance of the trained models, we calculated area under the operative characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV) or precision, negative categorisation value (NPV), accuracy, F-1 score and area under the precision-recall curve (AUPRC). Since the models were trained on class-imbalanced data, we focused on the AUPRC values to identify the best-performing models.

RESULTS

Descriptive summary

Of 70 680 self-reported PSE reports from November 2011 to October 2021 were extracted from PSRS. The PSE reports with unknown CFs were excluded from the study. In total, 53 899 PSE reports met the inclusion criterion. Online supplemental appendix B presents the frequency of the five CFs in our cohort. Not all CFs were included in this analysis; therefore, the percentages in online supplemental appendix B do not add up to one.

The PSE reports data were deidentified in terms of patient's name, date of birth, etc. However, the event narratives were not deidentified. The data are stored behind our Healthcare System's firewall, and it is not accessible to unauthorised users.

Information-rich terms and sentences

Table 1 presents the ngrams that were most influential when categorising the specific CFs. These ngrams were associated with high χ^2 and represented information-rich ngrams in each categorisation task.

The 25th percentile, median and 75th percentile of the number of sentences per PSE report were 4, 6, and 9, respectively. Communication/hand-off failure was the most sensitive and lapse/slip was the least sensitive to the changes in the number of selected information-rich ngrams.

Filtering out less relevant text changes the number of features to incorporate in a categorisation model. Figure 2 presents how the machine learning models' input dimension changed as we increased the number of selected information-rich ngrams. The input dimension would be almost 6700 if no sentence selection was applied. Unsurprisingly, the input dimension increased with the number of selected information-rich ngrams. Lapse/slip had the largest difference, while distractions/interruptions had the smallest difference.

Model comparison

The average number of sentences associated with each CF before and after incorporating information-rich sentence selection algorithm is presented in online supplemental appendix C. Moreover, the number of PSE reports grouped by sociotechnical CFs in training and testing data sets are included in online supplemental appendix D.

The XGBoost model performed best in categorising technology issues with AUPRC of 0.56 (figure 3). The

Table 1 Top five information-rich ngrams identified through χ^2 feature selection in each categorisation task

	Ngrams	χ^2 value
Communication/handoff failure	fall nurs order emerg depart depart	3786.7 2793.9 2462.0 1902.3 1853.6
Policy/procedure issue	min later fall polici glucomet glucose glucos run	1714.5 1474.9 1429.4 1408.1 1373.5
Technology issue	cgi field dose system medconnect	4917.1 1664.5 1646.7 1488.2 1425.2
Lapse/slip	dose pharmacy order pharmacist med	2158.7 1520.0 929.4 901.2 664.2
Distractions/interruptions	data entri vaccin error tablet medic	1686.5 1467.1 1452.1 1437.0 895.2

The words are stemmed.

FFNN model outperformed the other two models in categorising policy/procedure issues with AUPRC of 0.66. The XGBoost, elastic net and FFNN models achieved comparable performance in categorising communication/hand-off failure (AUPRCs=0.6, 0.58, 0.6), lapse/slip (AUPRCs=0.17, 0.15, 0.18) and distractions/interruptions (AUPRCs=0.24, 0.24, 0.22).

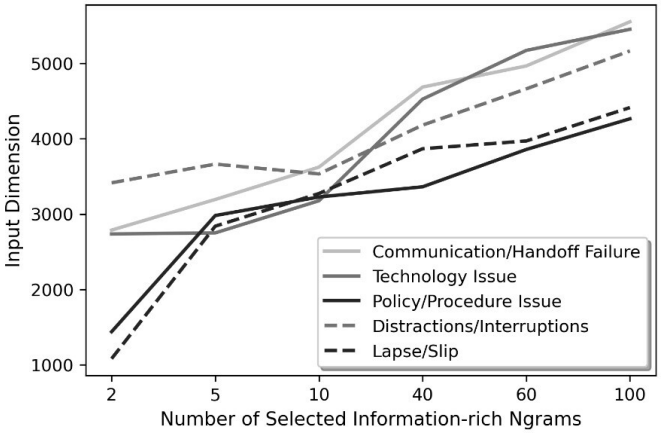

Figure 2 Input dimension vs the number of selected information-rich ngrams. Selecting the information-rich sentences, which include at least one of the selected information-rich ngrams, led to a new input feature dimension for contributing factor categorisation.

Besides AUPRC values, AUROC, sensitivity, specificity, PPV, NPV, F-1 score and accuracy were calculated for each model, and the results are included in the online supplemental appendix D. The elastic net model achieved highest AUROC (0.948) with baseline model in categorising distractions/interruption, highest sensitivity (0.879) with 60 information-rich ngrams in categorising distractions/interruptions, highest specificity (0.984) with two information-rich ngrams in categorising policy/procedure issue, highest PPV (0.692) with two information-rich ngrams in categorising technology issue, highest NPV (0.996) with 60 information-rich ngrams in categorising distractions/interruptions and highest accuracy (0.692) in categorising technology issue. The FFNN model obtained the highest F-1 score (0.962) with five information-rich ngrams in categorising policy/procedure issue.

DISCUSSION

This study used an NLP approach to identify information-rich sentences to categorise five CFs associated with sociotechnical errors. Automating the identification of CFs may help safety officers identify the CFs leading to safety issues in their organisations.

Utility of Information-rich sentences

Working with noise in the data is one of the challenges when dealing with free-text formatted input data. Filtering the input sentences and selecting more informative ones work as a solution to reduce the noise in the data when dealing with free-text categorisation of CFs. Finding a balance between removing the noise and keeping a sufficient number of features to have a well-trained model is a critical task.

The information-rich ngrams were selected according to their χ^2 values, indicating the association between ngrams and a CF. The most relevant term for communication/hand-off failure in our cohort was *fall* followed by *nurse* and *order*. *Glucose* readings were repeatedly identified as an important contributor to policy/procedure issues. Our data's selected information-rich ngrams for technology issues present the same trend through identifying *medconnect* and *system* as the most relevant ngrams for this CF. Besides, CGI, the stemmed form of Centigray (CGY), was the top information-rich ngram for technology issues. CGY is a measurement of absorbed radiation. This result suggests for technology-related CFs associated with radiation treatments such as with cyberknife radiation treatments. The ngrams with high χ^2 values associated with lapse/slip were *dose*, *pharmacy* and *order*. The pattern among these ngrams indicates that the medication-related tasks were more prone to be affected by this CF. Our analysis indicated that the data entry process was affected by distractions/interruptions as some of the information-rich ngrams for this CF were *data entry* and *error*. Box 1 presents examples of information-rich ngrams, which identified information-rich sentences in PSE reports.

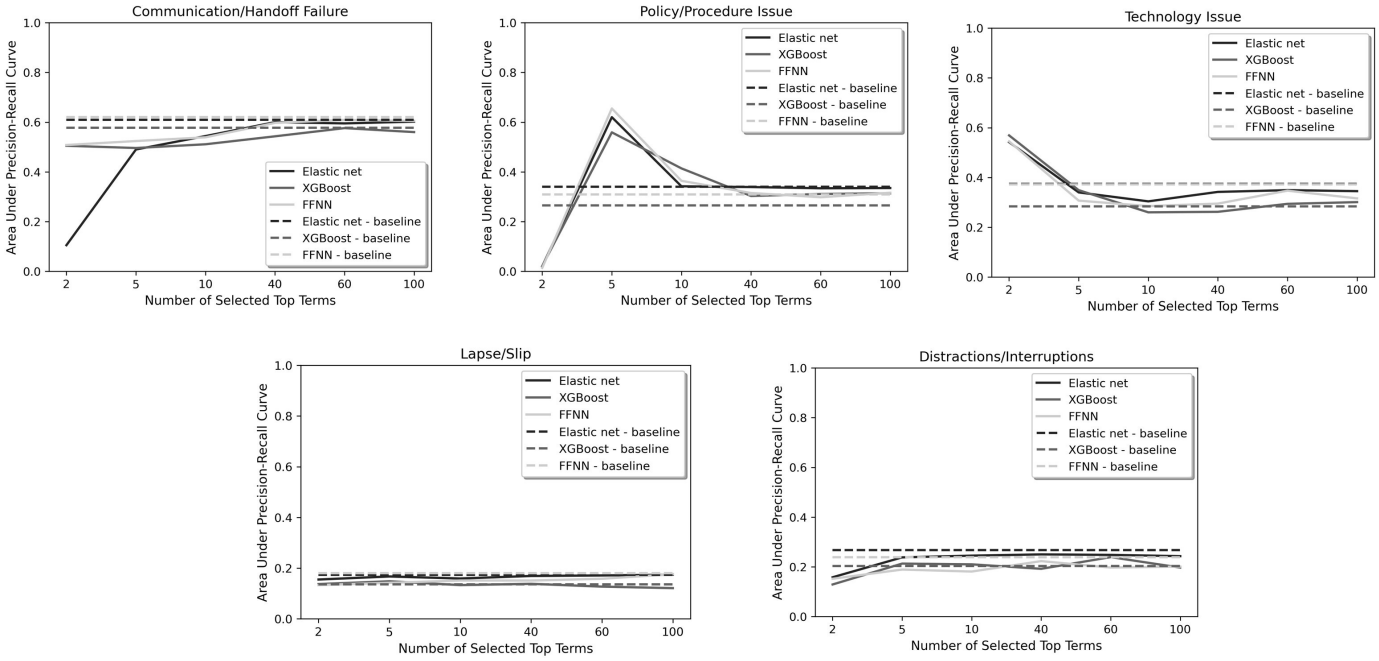

Figure 3 Area under precision-recall trends. For each of the five contributing factors, this figure presents how the value of AUPRC score changes as we only include the information-rich sentences in PSE reports containing the information-rich ngrams. AUPRC, area under the precision-recall curve; FFNN, feed-forward neural network; PSE, patient safety event.

We included unigrams and bigrams in the bag-of-words to calculate their χ^2 value. Information-rich unigrams were more common than bigrams, perhaps because unigrams contain more generalisable information. However, bigrams can convey more specific information, but they are susceptible to noise. Further investigation is needed to assess the effect of bigrams in the information-rich sentence selection algorithm.

Performance boosting

The proposed method proved its benefit by improving the results of categorising two CFs and achieving comparable results with the baseline models for three CFs in this analysis. Figure 2 shows a near-linear relationship between the number of input dimensions and the selected information-rich ngrams with higher χ^2 values. However, the AUPRC plots in figure 3 indicate that incorporating 2, 5 or 10 information-rich ngrams in selecting the information-rich sentences for the model training process may lead to better performance metrics. Thus, it can be inferred that the balance between removing noise and preserving features for an accurate model can be obtained by selecting the sentences containing the top 5 or 10 information-rich ngrams with the highest χ^2 values. This balance depends on the information embedded in the unstructured data.

Content depending

The elastic net model did not perform remarkably better than the neural network or ensemble model, implying that including the higher order interaction between the terms improved the categorisation performance. Identifying bigrams as information-rich terms also indicates the importance of the interactions among the features. FFNN and XGBoost models achieved comparable results. The difference between the three models' performance was negligible when categorising the CFs with lower prevalence, such as lapse/slip and distraction/interruption.

The information-rich sentence selection algorithm improved the performance of categorising two CFs, policy/procedure and technology issues. The performance improvement may also indicate that healthcare providers tend to be more consistent in their language to record safety events related to policy/procedure and technology issues. Distraction/interruption and lapse/slip had relatively smaller sample size compared with the other three CFs. However, the effect of information-rich sentence selection boosted the performance of distraction/interruption better. This is also an indication of having consistent language in recording safety incidents related to distraction/interruption.

Information-rich sentence selection is a data-driven approach; therefore, depending on the context of the unstructured input, the outcomes of using this approach could be different. For instance, the information-rich sentence selection approach improved the AUPRC values of all three machine learning categorisation models compared with baseline models in categorising policy/procedure and technology issue incorporating 5 and 2 information-rich ngrams, respectively. While the performance was not boosted for communication/hand-off failure, applying information-rich sentence selection approach using 100 information-rich terms led to similar results compared with using the entire PSE report for the categorisation task. Equivalently, using the approach

Box 1 Instances of information-rich sentence selection algorithm applied on patient safety event (PSE) reports from different contributing factor categories.

Contributing factor
Communication/hand-off failure.
PSE report (brief factual description)
Patient was taken to nuclear medicine via transport for a scheduled stress test. Once he got to NM, the test was cancelled. Patient had drunk coffee with his breakfast because there was no NPO order in place for the test.
Information-rich Ngram
Selected information-rich sentence
Patient had drunk coffee with his breakfast because there was no NPO order in place for the test.
Contributing factor
Policy/procedure issue
PSE report (brief factual description)
A glucose test was performed at (time stamp 1) on patient by (nures 1) with a result of 36 mg/dL. The test was performed at (time stamp 2) by (nurse 2) with a result of 139 mg/dL, which was 1 hour and 4 min later. The Hypoglycaemia Policy states that a patient with a glucose less than 40 mg/dL should be treated and a glucose run every 15 min until the glucose returns to 90 mg/dL.
Information-rich Ngram
Selected information-rich sentence
A glucose test was performed at (time stamp 1) on patient by (nures 1) with a result of 36 mg/dL. The Hypoglycaemia Policy states that a patient with a glucose less than 40 mg/dL should be treated and a glucose run every 15 min until the glucose returns to 90 mg/dL.
Contributing factor
Technology issue.
PSE report (brief factual description)
I was unable to gain access to pyxis. Rebooted system and tried several interventions but unsuccessful. ICU and ED called to report inability to gain access to pyxis. Carefusion called and stated that the database was disconnected from the system and unable to diagnose problem at this time. Instructed to call help line and high priority ticket initiated. Patient began seizing. Medication system down and unable to obtain ativan in the ED. Nurse had to physically go to the pharmacy to obtain medicine.
Information-rich Ngram
Selected information-rich sentence
Rebooted system and tried several interventions but unsuccessful. Carefusion called and stated that the database was disconnected from the system and unable to diagnose problem at this time. Medication system down and unable to obtain ativan in the ED.
Contributing factor
Lapse/slip.
PSE report (brief factual description)
Order for an HIV med entered on the wrong patient. The pharmacists did not quesiton why the patient was ordered for only one HIV medication. The doctor called one afternoon asking how did this mistake happen and not be caught. At that time, that is when the pharmacists was made aware of the mistake.
Information-rich Ngram
Pharmacist.
Selected information-rich sentence
The pharmacists did not quesiton why the patient was ordered for only one HIV medication. At that time, that is when the pharmacists was made aware of the mistake.

Continued

Box 1 Continued

Contributing factor
Distractions/interruptions.
PSE report (brief factual description)
Three prescriprions were e-scribed for one of our long-term patients here at store #N. All prescriptions were prepared and dispensed expediciously since our client was in a hurry to make his ride. All the medications were controlled except for one medication. The next day, we received a call from the doctor's office, which happens to be a first time doctor for this client, stating that one of the medications were to be dispensed at a later date on ((date)). Unfortunately, missed that date at data entry as I performed the data entry of the prescriptions. The team did contact the patient and informed him of the prescriber's specific directions in regards to that one prescription.
Information-rich Ngram
Selected information-rich sentence
Unfortunately, missed that date at data entry as I performed the data entry of the prescriptions.

with 100 information-rich ngrams for lapse/slip, and 60 information-rich ngrams for distraction/interruption did not improve the baseline AUPRC but achieved similar results with a filtered data and lower dimensional input for the machine learning models.

Limitations

Our study has limitations. First, our data came from a single health system and may reflect the specific language to the system. While PSE reports are recorded across all healthcare systems, the external application of our methods on data from other facilities may be biased. Second, our models were developed and evaluated based on a retrospective cohort; therefore, the performance may deteriorate when the method is applied to real-time data. Third, the five CFs included in this study are not the only CFs representing sociotechnical error. We explored the use of the NLP approach on only five sociotechnical CFs. This approach can be explored to categorising the rest of the sociotechnical CFs. Fourth, although we investigated the results and provided insights into the models' decision-making process, our study did not benefit from human factor expert input and critical analysis. Fifth, we selected TF-IDF, a widely known text feature extraction technique, and did not examine all text feature extraction methods (eg, YAKE!, rake, etc). Sixth, the CFs used in this study were assigned to PSE reports by reporters of safety incidents. The human-selected CFs could introduce some level of uncertainty to the labels. Seventh, we tested six values for the number of information-rich ngrams (ie, 2, 5, 10, 40, 60 and 100). Other values could be incorporated to measure the advantage of employing information-rich sentence selection algorithm. Finally, we excluded PSE reports, which did not have assigned CF, that can affect the performance of the models in near-real time applications.

Identifying and addressing CFs is critical for improving patient safety as these, often latent, themes are prevalent

across departments, event types and service lines. Being able to more readily identify CFs across departments, event types and service lines can provide patient safety leaders and healthcare systems awareness and insights to address safety events and hazards more at a system level.[37 38] This analysis is limited to what gets reported. While this is a useful start and one lens to understand CFs, a broader multiperspective approach is needed to understand all dimensions of CFs, including from the patient's perspective.[12 13] When analysing a large body of PSE reports and the associated CFs, it is essential to consider potential language bias and department bias (ie, using reports to 'blame and shame' other departments) in the recorded data.

CONCLUSION

We explored an NLP approach to categorise five socio-technical CFs in PSE reports. We demonstrated the utility of information-rich sentence selection in free-text categorisation. This approach can be used in near real-time to reduce the burden of manually extracting the factors influencing a patient's safety incident. Information such as patient feedback and complaints can be paired with the findings of this study to inform strategies around patient safety efforts and help teams make decisions.

Contributors Guarantor: AT, AF. Conceived study design: AT, AF. Contributed to data analysis: AT, AF. Contribute to visualisation: AT. Wrote the manuscript: AT, SS, ZMP, AF. Reviewed and edited manuscript: AT, SS, ZMP, AF.

Funding This work was supported by the National Institutes of Health grant number 1R01LM013309-01.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs

Azade Tabaie http://orcid.org/0000-0003-1869-5923
Zoe M Pruitt http://orcid.org/0000-0002-2749-9652
Allan Fong http://orcid.org/0000-0002-7550-1569

REFERENCES

1 Fong A. Realizing the power of text mining and natural language processing for analyzing patient safety event narratives: the challenges and path forward. *J Patient Saf* 2021;17:e834–6.
2 Archer S, Hull L, Soukup T, *et al*. Development of a theoretical framework of factors affecting patient safety incident reporting: a theoretical review of the literature. *BMJ Open* 2017;7:e017155.
3 Mitchell I, Schuster A, Smith K, *et al*. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'to err is human. *BMJ Qual Saf* 2016;25:92–9.
4 Macrae Carl. The problem with incident reporting. *BMJ Qual Saf* 2016;25:71–5.
5 Roehr B. US hospital incident reporting systems do not capture most adverse events. *BMJ* 2012;344(jan13 2):bmj.e386.
6 Flynn EA, Barker KN, Pepper GA, *et al*. Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *Am J Health Syst Pharm* 2002;59:436–46.
7 Pronovost PJ, Thompson DA, Holzmueller CG, *et al*. Toward learning from patient safety reporting systems. *Journal of Critical Care* 2006;21:305–15. 10.1016/j.jcrc.2006.07.001 Available: https://doi.org/10.1016/j.jcrc.2006.07.001
8 Holmström AR, Järvinen R, Laaksonen R, *et al*. Inter-Rater reliability of medication error classification in a voluntary patient safety incident reporting system Haipro in Finland. *Research in Social and Administrative Pharmacy* 2019;15:864–72.
9 Amaniyan S, Faldaas BO, Logan PA, *et al*. Learning from patient safety incidents in the emergency Department: a systematic review. *J Emerg Med* 2020;58:234–44.
10 Lacson R, Cochon L, Ip I, *et al*. Classifying safety events related to diagnostic imaging from a safety reporting system using a human factors framework. *Journal of the American College of Radiology* 2019;16:282–8.
11 Lawton R, McEachan RRC, Giles SJ, *et al*. Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. *BMJ Qual Saf* 2012;21:369–80.
12 Puthumana JS, Fong A, Blumenthal J, *et al*. Making patient safety event data actionable: understanding patient safety analyst needs. *J Patient Saf* 2021;17:e509–14.
13 Pronovost P, Morlock LL, Sexton B. Improving the value of patient safety reporting systems. In: *Advances in patient safety: New directions and alternative approaches. Vol 1. Assessment*. Rockville, MD: Agency for Healthcare Research and Quality, 2008.
14 Wood KE, Nash DB. Mandatory state-based error-reporting systems: current and future prospects. *Am J Med Qual* 2005;20:297–303.
15 Holden RJ, Carayon P, Gurses AP, *et al*. SEIPS 2.0: a human factors framework for studying and improving the work of Healthcare professionals and patients. *Ergonomics* 2013;56:1669–86.
16 Diller T, Helmrich G, Dunning S, *et al*. The human factors analysis classification system (HFACS) applied to health care. *Am J Med Qual* 2014;29:181–90.
17 Magrabi F, Ong MS, Runciman W, *et al*. An analysis of computer-related patient safety incidents to inform the development of a classification. *J Am Med Inform Assoc* 2010;17:663–70.
18 Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*. Stroudsburg, PA, USA, n.d.: 1532–43.
19 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their Compositionality. *Adv Neural Inf Process Syst* 2013;26.
20 Peters M, Neumann M, Zettlemoyer L, *et al*. Dissecting contextual word embeddings: architecture and representation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA, Brussels, Belgium.
21 Devlin J, Chang MW, Lee K, *et al*. Bert: pre-training of deep Bidirectional transformers for language understanding. 2018.
22 Ko Y, Park J, Seo J. Improving text Categorization using the importance of sentences. *Information Processing & Management* 2004;40:65–79.
23 Ogura Y, Kobayashi I. Text classification based on the latent topics of important sentences extracted by the Pagerank algorithm. In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. n.d.: 46–51.
24 Wright J, Lawton R, O'Hara J, *et al*. Assessing risk: a systematic review of factors contributing to patient safety incidents in hospital settings. improving patient safety through the involvement of patients: development and evaluation of novel interventions to engage patients in preventing patient safety incidents and protecting them against unintended harm. October 2016.

25 Chen YT, Chen MC. Using Chi-square Statistics to measure similarities for text Categorization. *Expert Systems with Applications* 2011;38:3085–90. 10.1016/j.eswa.2010.08.100 Available: https://doi.org/10.1016/j.eswa.2010.08.100
26 Fothergill R, Cook P, Baldwin T. Evaluating a topic Modelling approach to measuring corpus similarity. In: *Int Conf Lang Resour Eval*. 2016: 273–9.
27 Kilgarriff A. Using word frequency lists to measure corpus homogeneity and similarity between Corpora. *InFifth Workshop on Very Large Corpora* 1997.
28 Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical Taxonomy. 20, 1997.
29 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;67:301–20.
30 Lee SJ, Weinberg BD, Gore A, *et al*. A Scalable natural language processing for Inferring BT-RADS Categorization from unstructured brain magnetic resonance reports. *J Digit Imaging* 2020;33:1393–400.
31 Pfob A, Sidey-Gibbons C, Lee H-B, *et al*. Identification of breast cancer patients with pathologic complete response in the breast after Neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *Eur J Cancer* 2021;143:134–46.
32 Chen T, Guestrin C. Xgboost: A Scalable tree boosting system. *KDD* 2016;13:785–94.
33 Tabaie A, Orenstein EW, Nemati S, *et al*. Predicting presumed serious infection among hospitalized children on central venous lines with machine learning. *Comput Biol Med* 2021;132:104289.
34 Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* 1997;39:43–62.
35 Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014.
36 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *Jair* 2002;16:321–57.
37 Clarke JR. How a system for reporting medical errors can and cannot improve patient safety. *Am Surg* 2006;72:1088–91;
38 Howell AM, Burns EM, Hull L, *et al*. International recommendations for national patient safety incident reporting systems: an expert Delphi consensus-building process. *BMJ Qual Saf* 2017;26:150–63.