Journal of the American Medical Informatics Association, 27(10), 2020, 1502–1509
doi: 10.1093/jamia/ocaa082
Advance Access Publication Date: 18 June 2020
Research and Applications

Research and Applications

Can Unified Medical Language System–based semantic representation improve automated identification of patient safety incident reports by type and severity?

Ying Wang, Enrico Coiera, and Farah Magrabi *

Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

Corresponding Author: Farah Magrabi, PhD, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, North Ryde NSW 2113, Sydney, Australia; farah.magrabi@mq.edu.au

Received 10 February 2020; Revised 3 April 2020; Editorial Decision 21 2019; Accepted 27 April 2020

ABSTRACT

Objective: The study sought to evaluate the feasibility of using Unified Medical Language System (UMLS) semantic features for automated identification of reports about patient safety incidents by type and severity.
Materials and Methods: Binary support vector machine (SVM) classifier ensembles were trained and validated using balanced datasets of critical incident report texts (n_type = 2860, n_severity = 1160) collected from a statewide reporting system. Generalizability was evaluated on different and independent hospital-level reporting system. Concepts were extracted from report narratives using the UMLS Metathesaurus, and their relevance and frequency were used as semantic features. Performance was evaluated by F-score, Hamming loss, and exact match score and was compared with SVM ensembles using bag-of-words (BOW) features on 3 testing datasets (type/severity: n_benchmark = 286/116, n_original = 444/4837, n_independent =6000/5950).
Results: SVMs using semantic features met or outperformed those based on BOW features to identify 10 different incident types (F-score [semantics/BOW]: benchmark = 82.6%/69.4%; original = 77.9%/68.8%; independent = 78.0%/67.4%) and extreme-risk events (F-score [semantics/BOW]: benchmark = 87.3%/87.3%; original = 25.5%/19.8%; independent = 49.6%/52.7%). For incident type, the exact match score for semantic classifiers was consistently higher than BOW across all test datasets (exact match [semantics/BOW]: benchmark = 48.9%/39.9%; original = 57.9%/44.4%; independent = 59.5%/34.9%).
Discussion: BOW representations are not ideal for the automated identification of incident reports because they do not account for text semantics. UMLS semantic representations are likely to better capture information in report narratives, and thus may explain their superior performance.
Conclusions: UMLS-based semantic classifiers were effective in identifying incidents by type and extreme-risk events, providing better generalizability than classifiers using BOW.

Key words: UMLS, semantics, patient safety, incident reporting, supervised machine learning, text classification, natural language processing

INTRODUCTION

Health systems are experiencing a growing need to better harness their incident monitoring systems to improve patient safety. Incident monitoring is now widespread with the implementation of monitoring systems in most developed nations.[1] Healthcare professionals are routinely submitting reports about incidents or safety events in the delivery of care that could have resulted or did result in unnecessary harm to patients.[2] By retrospectively analyzing incident reports, health systems are striving to detect and manage emerging risks to patient safety.[3]

The timely detection and response to patient safety incidents is highly dependent on rapid analysis of reports by humans, which is a highly resource-intensive task. Patient safety experts typically triage incidents by their type and severity level for detailed analyses of unstructured free text data to identify contributing factors so that lessons can be learned and corrective action can be taken. Incident types are usually based on priority areas for safety and quality improvement (eg, falls, medications), whereas severity level is based on the consequences of incidents and the likelihood of recurrence.[4,5] For instance, hospital-level analyses could investigate common causes, contributing factors, and outcomes for specific incident types (eg, medications). Similarly, clusters of similar high-risk incidents could be identified at a health system level, indicating an emerging risk to patient safety. Take, for instance, an update to an order entry system that turns off the alerts for a high-risk medication across multiple sites. An incident at one site involving the medication may not be significant, whereas a cluster across multiple sites can facilitate early detection of the issue.

A major limitation of manual analysis is that it can no longer keep up with the growing volume of reports. For instance, 203 140 incidents were reported in the Australian state of New South Wales in 2018.[4] One way to improve the identification of incident clusters is to ask reporters to categorize incidents in a standardized manner using structures like the Agency for Healthcare Research and Quality Common Formats.[6] However, the labeling of incidents only works if reporters are knowledgeable about classification systems and are able to apply them consistently.[7] Studies have shown that the ratings provided by reporters are often inaccurate, and there is a high discordance in labels because health professionals may not have expertise in categorizing incidents.[7–9] Moreover, labeling may often be absent, incomplete, or delayed, reducing the ability to respond in near real time.[8]

Another—potentially more effective—way to improve the efficiency of incident analyses is to apply machine learning methods. Recent work in this area has demonstrated the feasibility of using supervised text classification methods to automatically sift through the large volumes of incident reporting data and identify specific clusters of reports for further detailed analysis by humans.[10] However, most studies have focused on using supervised methods for binary classification to identify a specific incident type. In reality, there are multiple incident types reflecting the breadth of problems in patient safety and little is known about the most optimal way to represent the free text from incident reports for automated identification. We recently demonstrated that a convolutional neural network with word embedding was effective in identifying incidents by type and severity, providing better generalizability than support vector machine (SVM) classifiers.[11] While a convolutional neural network is an elegant solution to distinguish multiple incident types, it does not support multiple labels.

Often an incident report can relate to more than 1 patient safety problem (ie, it can be assigned to multiple incident types).[1] For example, "Episode label A for patient X was placed incorrectly onto the specimen that belongs to patient Y," describes an error in patient identification that also relates to documentation. In another previous study to identify up to 2 labels for 10 incident types, we found that classifiers based on the bag-of-words (BOW) model failed when causes and consequences of incidents were implicitly described in reports.[12] This is because BOW does not take the semantic structure of language into consideration, and the ordering of words is also lost. For instance, an incident about a deteriorating patient was misidentified as a medications problem because there was a long list of medications and their doses in the incident narrative.[12] We also found that BOW classifiers were not effective in identifying rare classes of incidents (ie, making up < 2% of all incidents) in which the data available for training were limited (eg, infection, deteriorating patient).

Thus, in this study, we sought to evaluate the utility of semantic feature representation for automated identification incident reports by type and severity. We chose the Unified Medical Language System (UMLS) Metathesaurus, as it covers more than 2.9 million medical concepts from more than 150 source biomedical vocabularies.[13] The UMLS links alternative names for concepts and identifies useful relationships between concepts. The semantics in reports were represented by mapping incident narratives to UMLS concepts to train and validate (SVM) classifiers. For incident type, ensembles of binary classifier chains (ECCs) were used because they perform better than other multilabel classification algorithms, such as binary relevance,[14,15] and their chain structure can leverage the relationships among labels. We then compared the performance of this approach with BOW features from our previous work.[12,15]

MATERIALS AND METHODS

An overview of our approach comprising 4 main steps is shown in Figure 1.

1. *Data preprocessing*: Incident reports were collected from 2 separate reporting systems, the Advanced Incident Management System (AIMS) and Riskman.[4,5] The reports were labeled by patient safety experts to provide a gold standard label for experiments.
2. *Semantic features representation*: The narratives of reports were processed to generate semantic features.
3. *Training and validating classifiers*: Classifiers were trained and validated under cross-validation process to select optimal models based on performance. Here, the multilabel classification problem was transformed by training multiple basic binary classifier chains (CCs) (ie, developing a classifier identifying one type against all others while involving the relationships among incident types).[15] For severity level, we decomposed the problem into a series of binary classification problems. To distinguish 4 severity levels, 6 binary classifiers were trained using balanced datasets with one-vs-one (OvsO) strategy.[12]
4. *Testing classifiers:* Was undertaken with 3 testing datasets from the AIMS and Riskman systems. Incident type and severity level were examined separately. Model performance was evaluated and compared with BOW classifiers.[10] Each of these 4 steps is further detailed subsequently.

Step 1: Data preprocessing

Incident reporting systems

The reports were collected from 2 separate incident reporting systems: AIMS and Riskman.[4,16] AIMS has been used in Australia, New Zealand, South Africa, and the United States. In Australia, it has been used across the public hospital system in 4 of the 8 states and territories: New South Wales, Western Australia, South Australia, and the Northern Territory. In one Australian state, 137 522 incidents were reported to AIMS from January to December 2011. Of these, 6000 reports were selected using a random sampling approach.[5] Similarly, an independent set of 6000 reports were randomly selected from those that had been submitted to a hospital-level Riskman system from January 2005 to July 2012. The Riskman system is an independent tool used across the state of Victoria and a number of private hospitals.

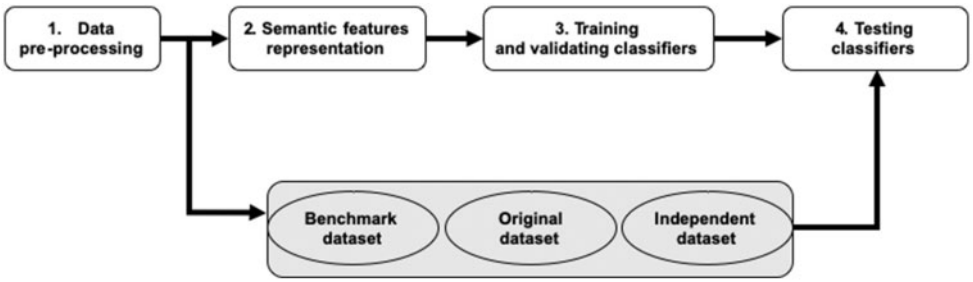

Figure 1. Flowchart to train, validate and test semantic classifiers using 3 testing datasets: benchmark, original, and independent.

Incident reports consist of a number of structured and free text fields to describe the event and its consequences (Box 1). Upon collection, any personally identifiable information was removed in accordance with jurisdictional privacy requirements (eg, name, date of birth). Only descriptive narratives were used in experiments including incident description, patient outcome, actions taken, prevention steps, investigation findings and results. All system-specific codes, punctuation, and nonalphanumerical characters were removed, and text was converted to lowercase. Ethical approval was obtained from university committees as well as a committee governing the hospital and state datasets.

BOX 1. THE BASIC ELEMENTS OF AN INCIDENT REPORT FROM THE AIMS AND RISKMAN SYSTEMS

Report format	Structured	Free text
Basic element	incident ID date and time • incident type(s) • severity access code	description of incident actions taken • preventative steps • patient outcome investigation findings and results

Labeling reports

Given the inconsistency of reporter labels,[9] 3 experts in the classification of incidents reviewed and validated labels for all reports based on the international classification for patient safety.[5] The reports were split among the 3 experts, and the labels they provided were used as "gold standard" for training and testing the performance of classifiers. The reports were classified into 20 incident types, and this study focused on 10 of these types that have been recognized as priority areas for safety and quality improvement (Table 1).[2,5,12] Interrater reliability for determining incident types was Cohen's kappa = 0.93 ($P < .001$; 95% confidence interval, 0.9301-0.9319). To cover the whole dataset, an "Others" set was created using a random sampling approach to ensure representativeness of 10 other unrelated incident types—see Appendix B in in Wang et al.[12]

The seriousness of an incident was rated by an internationally accepted rating system called the severity assessment codes (SACs), developed by the U.S. Veterans Administration.[17] Given the severity of an incident and the likelihood of recurrence, 4 risk ratings (i, extreme; ii, high; iii, medium; and iv, low) were used.[18] The gold standard was based on assignment of SAC ratings for every report by local patient safety managers who had received training in assessing incident seriousness and were familiar with the nature of incidents and their consequences.

Generating training and testing subsets

Of the 6000 AIMS reports, we used 260 that were randomly selected from each incident type and 290, also randomly selected, for each SAC level to create a balanced dataset (Table 1). The sample sizes were based on previous studies.[12] The balanced dataset was further divided into training (80%), validation (10%), and testing (10%) subsets under a 10-fold subsampling cross-validation process. The training and validation subsets were used to identify the most effective classifiers, and the testing subset (benchmark) was applied to generate benchmark results.

To evaluate applicability in real-world conditions, classifiers were further tested on imbalanced (ie, "stratified") datasets from AIMS (original). These stratified datasets were randomly selected from the remaining AIMS reports based on the real-world ratio incidents by type and severity (Table 1). To examine generalizability to an independent incident reporting system, classifiers were tested on a stratified Riskman dataset (independent).

Step 2: Semantic features representation

We developed a bag-of-concepts model to extract unique concepts as semantic features from incident narratives. This approach is similar to a BOW model extracting unique words but maps tokens, terms, and words from incident reports into medical concepts from the UMLS. To identify unique concepts, a software tool called MetaMap was used to annotate the UMLS concepts and semantic types.[19] The bag-of-concepts model was represented as an unordered collection of concepts in which each concept was used as a feature. Three feature engineering approaches were adopted to represent the occurrence of concepts: binary count, term frequency (TF), and TF–inverse document frequency (TF-IDF).[20]

Step 3: Training and validating classifiers

Ensemble strategy

The classification problem was decomposed into a number of binary classification problems.[12,15] There are 2 traditional ensemble strategies to pool decisions from each base binary classifier: one-vs-all (OvsA) and OvsO.[21] OvsA divides an n class problem into n binary problems by training classifiers to distinguish one class from all other classes.[21,22] It was used in identifying incident types as it considered all possible label combinations, suiting the design of learning label relationships.[15,23] The OvsO strategy transforms an n class problem into $n*(n-1)/2$ binary problems by involving all possible combinations between pairs of classes. Given its good performance

Table 1. Composition of balanced and stratified datasets for training and testing semantic classifiers

	Benchmark (balanced AIMS)		Original (stratified AIMS)			Independent (stratified Riskman)		
	N1	N2	N1	N2	%	N1	N2	%
Incident type								
Falls	260	261	90	91	20	872	939	15
Medications	260	304	68	74	15	1053	1217	18
Pressure injury	260	264	37	38	8	190	197	3
Aggression	260	271	49	57	11	487	541	8
Documentation	260	589	26	67	6	252	809	4
Blood product[a]	260	273	5	6	1	59	70	1
Patient identification[a]	260	337	7	8	2	86	117	1
Infection[a]	260	274	6	6	1	22	35	<1
Clinical handover[a]	260	301	7	8	2	87	101	1
Deteriorating patient[a]	260	264	1	2	<1	14	21	<1
Others	260	689	148	173	33	2,878	4039	48
Total	2860	3827	444	530		6000	8086	
Severity level								
SAC1	• 290		• 25		<1	• 23		<1
SAC2	• 290		• 95		2	• 105		2
SAC3	• 290		• 2198		45	• 2609		44
SAC4	• 290		• 2519		52	• 3213		54
Total	• 1160		• 4837			• 5950		

The same data was used to train bag-of-words classifiers.

N1 is the number of reports based on primary labels. N2 is the number of reports considering 2 labels, and % is based on primary label alone.

AIMS: Advanced Incident Management System; SAC: severity assessment code.

[a]Rare incident type (ie, <2%).

to identify severity levels in our previous work,[12] the OvsO strategy was adopted.

Base binary classifiers

SVM with radial basis function (RBF) kernel was chosen, as it has been shown to be effective for smaller datasets with a large feature space.[24,25] Furthermore, SVM outperformed other binary classifiers in our previous work, such as a logistic regression model.[12,15]

We used an ensemble of SVM classifiers with semantic features.[15,26] For incident type, SVM classifiers were linked along a chain where each classifier dealt with a binary problem. Classifier chains were different from the common binary classifiers because their feature space was extended with binary values to indicate the relationships between incident types that appeared in training data.[26] The order of the chain itself was determined by the order of the label variables. In the extended feature space, binary values of labels only indicated which of previous labels were assigned to reports.[14] To reduce the influence of label order, incident types were randomly reordered and training was repeated. Final decisions were based on the predictions of each ECC. As classification performance often increases with ECC size used in training, ECCs may become unnecessarily large and impose a high computational cost. We examined the optimal size for ECC combinations by varying the number of ensembles from 3 to 40. A learning curve was used to find the optimal ECC by trading off computational cost against classifier accuracy.[26] The feature space for severity level involved the sematic features alone. This is because severity level involves a single label or risk group while there may be multiple incident types.

Base classifiers were trained, validated, and tested on balanced datasets using 3 semantic feature representations: binary count, TF, and TF-IDF. A 10-fold subsampling cross-validation method was applied to optimize the classifier parameters. The parameters from classifiers achieving the highest F-score were adopted for testing.

Step 4: Testing classifiers

In group decision making, majority vote is an efficient method and has been commonly accepted in classification ensembles.[25] To identify incident type, the final decisions were made by averaging multiple predictions from individual CCs. Up to 2 labels were predicted if the averaged classification probabilities exceeded a predefined threshold. For severity level, each base classifier voted and the final prediction was based on the level with the most votes.[27]

For incident type, overall performance was evaluated using example-based and label-based measures.[15] Example-based measures evaluate the difference of the true and the predicted sets of labels over all testing reports. Label-based measures consider additional degrees of freedom with multilabels and evaluate performance separately for each incident type, and then average the performance over all types. We used 6 commonly example-based measures, including Hamming loss, accuracy, exact match score, precision, recall, and F-score.[28] Given 2 labels per report, the performance for individual types was evaluated based on *OR* logic when matching the predicted and true set of labels. Label-based measures included macro-averaged and micro-averaged precision, recall, and F-score.[12,15] We also examined the top 20 UMLS concepts from each incident type to identify concepts that were common across the 10 types of incidents, as well as those that were unique to each incident type.

For severity level, overall performance was evaluated using micro-averaged measures of precision, recall, and F-score based on the cumulative number of true positives, true negatives, false positives, and false negatives per type.[12,28] When identifying a specific severity level, the F-score, precision, and recall measures were evaluated per level. Individual measures were calculated based on the probability that a specific severity level was classified as such (eg, % of extreme-risk incidents correctly identified among the test set for SAC1).

Table 2. The most effective classifiers for incident type using semantic features compared with BOW

Classification studies	Semantics	BOW[15]
Ensemble strategy	OvsA	OvsA
Ensemble size	12 ECCs	6 ECCs
Feature extraction	Bag of concepts	BOW
Feature space representation	TF-IDF + label associations	Binary count + label associations
Base classifier	SVM RBF kernel	SVM RBF kernel
Group decision making	Voting	Voting
Average F-score		
Benchmark dataset, %	82.6	69.4
Original dataset, %	77.9	68.8
Independent dataset, %	78.0	67.4

BOW: bag-of-words; ECCs: ensemble of binary classifier chains; OvsA: one vs all; RBF: radial basis function; SVM: support vector machine; TF-IDF: term frequency–inverse document frequency.

Table 3. Overall classification performance of semantic features in identifying incident type compared with BOW

	Benchmark		Original		Independent	
Feature representation	Semantics	BOW	Semantics	BOW	Semantics	BOW
Example-based measures						
Hamming loss	3.9	7.8	3.7	7.2	3.9	8.1
Accuracy	82.2	64.4	75.6	68.0	75.4	61.7
Precision	84.7	70.6	77.3	72.9	78.1	66.4
Recall	84.4	77.1	76.7	76.6	76.8	67.4
F-score	84.6	73.7	77.0	74.7	77.4	66.9
Exact match	48.9	39.9	57.9	44.4	59.5	34.9
Label-based measures						
Macro-precision	86.7	69.7	77.7	52.4	70.2	54.2
Macro-recall	91.1	79.0	79.9	77.4	78.2	70.3
Macro–F-score	87.9	73.7	77.1	59.2	73.6	58.8
Micro-precision	82.8	67.1	78.0	67.1	78.9	68.7
Micro-recall	82.4	71.9	77.9	70.7	77.1	66.1
Micro–F-score	82.6	69.4	77.9	68.8	78.0	67.4

Values are %.
BOW: bag-of-words.

RESULTS

Overall performance to identify incident type

For incident type, the most effective classifier was an ensemble of 12 binary classifier chains of SVM RBF kernel with TF-IDF UMLS-based semantic feature representation, achieving an averaged F-score of 82.6% on the benchmark, 77.9% on the original, and 78.0%, on the independent datasets (Table 2).

Compared with BOW, the overall performance of semantic classifiers was superior (Table 3). When identifying multiple incident types to which reports could be assigned, semantic classifiers achieved better exact match scores, improving overall performance on all 3 testing datasets. Hamming loss dramatically dropped from 7.2 to 3.7 on the original dataset and decreased from 8.1 to 3.9 on the independent dataset. By label-based measures, both semantic and BOW classifiers performed consistently on the original and independent datasets. With the exception of macro-recall, all label-based measures were higher.

Identifying individual incident types

On the benchmark dataset, semantic classifiers outperformed BOW when identifying all incident types except falls and deteriorating patient, which remained steady at 88.1% and 87.5%, respectively (Table 4). With the original dataset, F-score improved, ranging from 64% for patient identification to 92.4% for falls. On the independent dataset, F-scores improved for all incident types except documentation, which remained steady at 75.2% (BOW: 75%), and for deteriorating patient, in which it worsened to 48.8% (BOW: 54.6%).

When identifying the rare incident types of blood products, patient identification, infection, and clinical handover, performance of the semantic classifiers improved dramatically on the original and independent datasets (Table 4). For instance, the F-score for blood products increased from 44.4% to 75% on the original; and from 54.9% to 76.9% on the independent. Patient identification incidents were frequently misclassified by BOW as documentation and others (F-score: original = 34.0%; independent = 56.4%). However, semantic classifiers improved this performance to 64.0% and 70.5%. For infection, F-scores were also improved (original: 46.2% to 80%; independent 30.8% to 63.1%). For clinical handover, F-scores for the semantic classifiers ranged from 50.4% to 51.6%, compared with from 32% to 34.5% with BOW.[15] The exception was deteriorating patient, in which F-score improved on the original dataset (from 44.4% to 66.7%), but was worse on the independent dataset (from 54.6% to 48.8%).

Mapping between UMLS concepts and incident types

We reviewed the top 20 concepts from each incident type (see Supplementary Appendix). Not surprisingly, the UMLS concept that was common across all types was patients. Other concepts that were common across 8-9 incident types were ward, result, present, and time (see Supplementary Appendix Table 5). Physicians, nurses, and notification were common across 6 incident types. Concepts that were unique to individual incident types generally related to specific clinical tasks and procedures (see Supplementary Table 6). For instance, the concepts dosage, prescribed, pharmacy, and pharmacist infusion procedures were unique to medications, whereas secluding patient, aggressive behavior, and escort were unique to aggression incidents. For blood products, unique concepts included peripheral blood, blood transfusion, blood product, blood bank, and departments such as laboratory and pathology. For documentation, unique concepts included signature and electronic health records, whereas

Table 4. F-score for identifying incident types and severity level using semantic features compared with BOW

	Benchmark		Original		Independent	
Feature representation	Semantics	BOW[12,15]	Semantics	BOW[12,15]	Semantics	BOW[12,15]
Incident type						
Falls	88.1	88.1	92.4	89.7	90.3	81.0
Medications	78.5	67.4	87.1	76.3	85.1	75.2
Pressure injury	96.4	91.5	91.6	85.4	85.7	81.4
Aggression	93.1	74.0	82.6	69.1	79.3	63.8
Documentation	71.0	62.2	78.2	61.2	75.2	75.0
Blood products[a]	96.6	71.1	75.0	44.4	76.9	54.9
Patient identification[a]	77.2	66.0	64.0	34.0	70.5	56.4
Infection[a]	93.8	81.1	80.0	46.2	63.1	30.8
Clinical handover[a]	77.0	62.5	51.6	32.0	50.4	34.5
Deteriorating patient[a]	87.5	89.7	66.7	44.4	48.8	54.6
Others	75.8	60.0	78.6	67.9	83.4	69.4
Severity level						
Average F-score	71.6	62.9	42.2	50.1	49.6	52.7
SAC1	87.3	87.3	25.5	19.8	50.7	12.5
SAC2	69.8	49.0	8.4	12.3	11.9	12.0
SAC3	75.3	49.1	42.1	42.6	58.4	48.3
SAC4	60.0	64.0	52.2	61.8	39.3	60.0

Values are %.
BOW: bag-of-words; SAC: severity assessment code.
[a]Rare incident type (ie, <2%).

transfer of care and communication were unique to clinical handover.

Identifying severity level

For severity level, OvsO ensembles of SVM RBF kernel with TF-IDF feature representation was the most effective combination. While the micro-averaged F-score (71.6%) was higher than BOW on the benchmark, it was lower on the original (42.2%) and independent (49.6%) datasets (Table 4). For SAC1, performance of the semantic classifiers was consistent with BOW on the benchmark (F-score-= 87.3%) but markedly improved on the original (F-score up from 19.8% to 25.5%) and independent (F-score up from 12.5% to 50.7%) datasets. For SAC2 and SAC3, semantic classifiers performed better than BOW on the benchmark dataset (SAC2: F-score up from 49.0% to 69.8%; SAC3: F-score up from 49.1% to 75.3.8%), but this improvement did not carry over to the original and independent datasets. For SAC4, F-score was consistently lower than BOW across all testing datasets.

DISCUSSION

Main findings and implications

We evaluated the feasibility of a UMLS-based semantic feature representation by comparing it with BOW features and found that it can enhance the performance of SVM classifiers in identifying patient safety incidents by type and severity, particularly high-risk SAC1 events. To the best of our knowledge, no previous studies have compared these 2 approaches to represent the free text from incident reports for automated identification.[10] In terms of overall performance, semantic feature representation improved generalizability, showing consistent performance from the original (F-score: 77.9%) to independent (F-score: 78.0%) datasets. Identification of primary and secondary labels was also improved; semantic classifiers achieved exact match scores of 57.9% and 59.5% on the original and independent datasets (BOW: 44.4%/34.9%).

For the different incident types, semantic classifiers improved identification of 8 types (ie, medications, pressure injury, aggression, documentation, blood products, patient identification, infection, and clinical handover). F-score was above 75% for 7 of 10 types in the original dataset. These 7 types made up 95% of all reported incidents (Table 1). There was a marked improvement in identifying rare incident types and extreme-risk events, indicating that UMLS concepts better capture information in report narratives compared with BOW and may thus be an effective strategy to address the highly unbalanced distribution of incident classes, particularly for rare incident types and high-risk incidents, which make up <2% of incidents. These results show that semantic classifiers have the potential to be applied in real-world settings. When human resources are lacking, automated methods can reduce effort by supporting the first step in incident analysis. Rapidly screening which incidents require immediate attention, and grouping incidents by type, allows human efforts to focus on a much smaller volume of incident reports. Automated methods are currently meant to be used in conjunction with expert review and are not intended to replace human input.

While the feasibility of automated incident identification has been demonstrated in controlled experiments, further work is required to understand the extent to which these methods can be generalized and how best they can be adapted to new settings to help identify clusters of potentially related reports with an underlying common cause. There is a need to trial text classifiers in real-world settings and to shift research from its sole focus on algorithmic performance to studying, in parallel, strategies for successful implementation and impact of these models on quality improvement initiatives and patient safety outcomes.[29] Any translation, sharing, and reuse of text classifiers should seek to leverage the growing use of common data formats and platforms for machine-executable models.[30]

Feature representation

We noted that the most effective feature representation for semantic classifiers was TF-IDF and binary count for BOW. TF-IDF reflects how important a word is to a report from all training documents, and its term-weighting scheme decreases the weight for commonly used concepts and increases the weight for infrequent concepts (ie, those might be specific to individual rare types). This might explain why the semantic features represented more distinctive concepts from rare classes and achieved better performance. Many clinical tasks and procedures unique to specific incident types were captured quite well by the UMLS concepts (Supplementary Table 6), contributing to better performance. However, these concepts did not focus on patient safety. To improve classification performance, more broadly applicable concepts for specific incident types are required. Better UMLS coverage of essential patient safety terminologies would be helpful for large-scale analysis of patient safety incidents. Another approach for incident types not adequately covered is to combine UMLS concepts with BOW features. This is an area that requires further investigation.

Identifying severity level

For severity level, semantic features improved the identification of SAC1 incidents in stratified datasets. On the independent dataset, F-score increased dramatically from 12.5% (BOW) to 50.7% (semantics). However, there was no consistent improvement for SAC2 and SAC3 incidents. SAC4 incidents were more likely misclassified as SAC3 and SAC2 by the semantic classifiers, achieving a lower F-score. As SAC4 is a common class, the average F-score across all levels was lowered. SAC4 was mostly misclassified as SAC3. We observed that the reports from SAC3 and SAC4 covered all 10 incident types, making it hard for them to be distinguished from each other, as the training reports from both severity levels generated similar semantic feature spaces. In addition, the involvement of 10 incident types resulted in a sparse semantic feature space to cover all the concepts representing the different clinical tasks and procedures. In general, sparse input requires more training samples to achieve reliable prediction. However, we did not have access to a larger training dataset. One possible solution is to apply feature engineering, such as feature selection, to obtain more informative features but with lower dimensionality. The semantic feature spaces from the medium and low severity levels of SAC3 and SAC4 were overlapped. The incidents from these 2 levels covered similar types of incidents, leading to the harder boundary in feature space for distinguishing them.

Limitations and future work

First, we used datasets from 2 independent reporting systems, but both came from the same Australian state. Therefore, the semantic classifiers may not be generalizable to other jurisdictions and regions using different reporting, linguistic styles, and terminology. Second, the training datasets were balanced. Given the class imbalance problem with incident types in real-world settings, a stratified training dataset may be more desirable when large training datasets are available. Last, to identify the different incident types, we used the same threshold to make final decisions when evaluating results from binary classifiers. We observed that classifiers identified common types with more confident probabilities (eg, most falls achieved a probability higher than 0.9), but the probabilities for rare classes were relatively lower (eg, infection ~0.5). This indicates that the same threshold for common and rare types may not be optimal, as it may result in some reports belonging to rare types being missed. Further experiments are required to investigate optimal thresholds for common and rare incident types.

CONCLUSION

Our experiments showed that semantic classifiers using UMLS concepts outperformed classifiers based on BOW features for identifying most incident types and extreme risk events. Semantic classifiers demonstrated better generalizability on the independent dataset by combing TF-IDF feature representation. Analysis of concepts unique to specific incident types indicated that semantic representation provides more distinct features to specify clinical tasks and procedures and may thus improve identification of incident types. Further work is required to improve identification of lower severity levels where decision boundaries are overlapped.

FUNDING

This research is supported in part by grants from the Australian National Health and Medical Research Council, project grant APP1022964 (FM); and the Centre for Research Excellence in Digital Health, grant 1134919 (EC and FM). The funding source did not play any role in the study design; collection, analysis, and interpretation of data; writing of the report; or decision to submit the article for publication.

AUTHOR CONTRIBUTIONS

EC, FM, and YW conceptualized the study. YW designed and implemented the training and evaluation of classification model, and is responsible for the integrity of the work. YW and FM drafted the article. All authors participated in writing and revising the article. All aspects of the study (including design; collection, analysis and interpretation of data; writing of the report; and decision to publish) were led by the authors. All authors read and approved the final manuscript.

ETHICS APPROVAL

Ethical approvals were obtained from committees of Macquarie University, the University of New South Wales as well as the committees governing the hospital and the state datasets.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We thank Bronwyn Shumack, Katrina Pappas, and Diana Arachi for assisting with the extraction of the incident reports. We also thank Anita Deakin, Alison Agers, and Sara Suffolk for their assistance with labeling reports.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Ramirez E, Martin A, Villan Y, *et al* Effectiveness and limitations of an incident-reporting system analyzed by local clinical safety leaders in a tertiary hospital: prospective evaluation through real-time observations of patient safety incidents. *Medicine (Baltimore)* 2018; 97 (38): e12509.

2. Runciman W, Hibbert P, Thomson R, *et al* Towards an international classification for patient safety: key concepts and terms. *Int J Qual Health Care* 2009; 21 (1): 18–26.
3. Bolsin SN, Colson M, Patrick A, *et al* Critical incident reporting and learning. *Br J Anaesth* 2010; 105 (5): 698.
4. Clinical Incident Management in the NSW Public Health System. Clinical Excellence Commission (CEC) and NSW Department of Health. Biannual incident report. http://www.cec.health.nsw.gov.au/clinical-incident-management Accessed May 19, 2020.
5. Runciman WB, Williamson JA, Deakin A, *et al* An integrated framework for safety, quality and risk management: an information and incident management system based on a universal patient safety classification. *Qual Saf Health Care* 2006; 15 (suppl 1): i82–90.
6. Agency for Healthcare Research and Quality. Patient Safety Organization (PSO) Program. Common formats. 2016. https://www.pso.ahrq.gov/common/ Accessed May 19, 2020.
7. Williams SD, Ashcroft DM. Medication errors: how reliable are the severity ratings reported to the national reporting and learning system? *Int J Qual Health Care* 2009; 21 (5): 316–20.
8. Gong Y. Data consistency in a voluntary medical incident reporting system. *J Med Syst* 2011; 35 (4): 609–15.
9. Haines TP, Massey B, Varghese P, *et al* Inconsistency in classification and reporting of in-hospital falls. *J Am Geriatr Soc* 2009; 57 (3): 517–23.
10. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform* 2019; 132: 103971.
11. Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. *J Am Med Inform Assn* 2019; 26 (12): 1600–8.
12. Wang Y, Coiera E, Runciman W, *et al* Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *BMC Med Inform Decis Mak* 2017; 17: 84.
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32: D267–70.
14. Liang C, Gong Y. Automated classification of multi-labeled patient safety reports: a shift from quantity to quality measure. *Stud Health Technol Inform* 2017; 245: 1070–4.
15. Wang Y, Coiera E, Runciman W, *et al* Automating the identification of patient safety incident reports using multi-label classification. *Stud Health Technol Inform* 2017; 245: 609–13.
16. Victoria State Goverment. Incident reporting in Victoria. Clinical risk management. https://www.health.vic.gov.au/clinrisk/vhims/index.htm Accessed January 31, 2020.
17. Runciman B, Walton M. *Safety and Ethics in Healthcare: A Guide to Getting it Right*. Farnham, United Kingdom: Ashgate; 2007.
18. Bagian JP, Lee C, Gosbee J, *et al* Developing and deploying a patient safety program in a large health care delivery system: you can't fix what you don't know about. *Jt Comm J Qual Improv* 2001; 27 (10): 522–32.
19. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17 (3): 229–36.
20. Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cybern* 2010; 1 (1–4): 43–52.
21. Dietterich TG. Ensemble methods in machine learning. In: *Lecture Notes in Computer Science: Multiple Classifier Systems*. New York, NY: Springer; 2000; 1857: 1–15.
22. Galar M, Fernandez A, Barrenechea E, *et al* An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. *Pattern Recogn* 2011; 44 (8): 1761–76.
23. Madjarov G, Kocev D, Gjorgjevikj D, *et al* An extensive experimental comparison of methods for multi-label learning. *Pattern Recogn* 2012; 45 (9): 3084–104.
24. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J Mach Learn Res* 2002; 2 (2): 265–92.
25. Sun MH. A multi-class support vector machine: theory and model. *Int J Inf Technol Decis Mak* 2013; 12 (06): 1175–99.
26. Read J, Pfahringer B, Holmes G, *et al* Classifier chains for multi-label classification. machine learning and knowledge discovery in databases. *Mach Learn* 2009; 5782: 254–69.
27. Black D. On the rationale of group decision-making. *J Polit Econ* 1948; 56 (1): 23–34.
28. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inform Process Manag* 2009; 45 (4): 427–37.
29. Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res* 2019; 21 (11): e16323.
30. Friedman CP, Flynn AJ. Computable knowledge: an imperative for learning health systems. *Learn Health Syst* 2019; 3 (4): e10203.