Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection

Taxiarchis Botsis,[1,2] Michael D Nguyen,[1] Emily Jane Woo,[1] Marianthi Markatou,[3,4] Robert Ball[1]

► Additional materials are published online only. To view these files please visit the journal online (www.jamia.org).

[1]Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Rockville, Maryland, USA
[2]Department of Computer Science, University of Tromsø, Tromsø, Norway
[3]Department of Statistical Sciences, Cornell University, New York, New York, USA
[4]IBM T.J. Watson Research Center, Hawthorne, New York, New York, USA

Correspondence to
Dr Taxiarchis Botsis, Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Woodmont Office Complex 1, Rm 306N, 1401 Rockville Pike, Rockville, Maryland 20852, USA;
taxiarchis.botsis@fda.hhs.gov

MM and RB have contributed as senior authors to this work.

Received 2 November 2010
Accepted 23 May 2011
Published Online First 27 June 2011

ABSTRACT

Objective The US Vaccine Adverse Event Reporting System (VAERS) collects spontaneous reports of adverse events following vaccination. Medical officers review the reports and often apply standardized case definitions, such as those developed by the Brighton Collaboration. Our objective was to demonstrate a multi-level text mining approach for automated text classification of VAERS reports that could potentially reduce human workload.

Design We selected 6034 VAERS reports for H1N1 vaccine that were classified by medical officers as potentially positive (N_{pos}=237) or negative for anaphylaxis. We created a categorized corpus of text files that included the class label and the symptom text field of each report. A validation set of 1100 labeled text files was also used. Text mining techniques were applied to extract three feature sets for important keywords, low- and high-level patterns. A rule-based classifier processed the high-level feature representation, while several machine learning classifiers were trained for the remaining two feature representations.

Measurements Classifiers' performance was evaluated by macro-averaging recall, precision, and F-measure, and Friedman's test; misclassification error rate analysis was also performed.

Results Rule-based classifier, boosted trees, and weighted support vector machines performed well in terms of macro-recall, however at the expense of a higher mean misclassification error rate. The rule-based classifier performed very well in terms of average sensitivity and specificity (79.05% and 94.80%, respectively).

Conclusion Our validated results showed the possibility of developing effective medical text classifiers for VAERS reports by combining text mining with informative feature selection; this strategy has the potential to reduce reviewer workload considerably.

INTRODUCTION

Biomedical research is often confronted with large data sets containing vast amounts of free text that have remained largely untapped sources of information. The analysis of these data sets poses unique challenges, particularly when the goal is knowledge discovery and real-time surveillance.[1] Spontaneous reporting systems (SRSs), such as the US Vaccine Adverse Event Reporting System (VAERS), encounter this issue.[2] When extraordinary events occur, such as the H1N1 pandemic, routine methods of safety surveillance struggle to produce timely results due to the resource-intensive nature of manual review. Consequently, there is an urgent need to develop alternative approaches that facilitate efficient report review and identification of safety issues resulting from the administration of vaccines. Text classification (TC) provides an alternative and more efficient process by distinguishing the most relevant information from adverse event (AE) reports.

Medical TC is the process of assigning labels to a span of text (sentence, paragraph, or document) using trained or rule-based classifiers,[3–10] or both.[11 12] The utilization of natural language processing (NLP) techniques may provide better classification results through improvements in text exploration. However, according to Cohen and Hersh, TC should be placed closer to the text mining (TM) field than the full-blown NLP field.[13] TM and NLP techniques have been used before to identify AEs in electronic health records (EHRs)[14–17]; however, the issue of a complete surveillance system that could be generalized has not been addressed yet.

Safety surveillance in VAERS (and other SRSs) has two main purposes. The first purpose is monitoring known adverse effects for unusual features or increases in reporting rate (ie, number of reports/number of doses) while looking for potential associations with new products (eg, H1N1 vaccine) or new demographic groups. The second purpose is looking for unexpected AEs by identifying unusual patterns. In the first case, we are more interested in the identification of the actual adverse cases, while in the second case we primarily need to know whether the identified patterns represent 'real' conditions in terms of clinical syndromes.

Here, we present a multi-level TM approach that was applied to a group of VAERS reports involving the AE of anaphylaxis for text classification purposes. This investigation of TM for anaphylaxis could serve as a model for TM in the first purpose of safety surveillance in VAERS and could also serve as the basis to generalize our work to other AEs that are acute, serious, and occur in close temporal proximity to the vaccination. Our scope was not to present a fully developed system for AE identification but rather to study patterns in the narrative of VAERS reports that are used to identify a known adverse effect. The strength of our study lies in demonstrating the feasibility of using TM on large SRS databases to exploit the information content of a new data source, other than EHRs or clinical

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

trials data, for adverse event identification as well as in saving time and human resources.

BACKGROUND

Vaccine Adverse Event Reporting System

The Vaccine Adverse Event Reporting System is a passive surveillance repository that monitors the number and type of AEs that occur after the administration of vaccines licensed for use in the USA.[18] VAERS contains both structured (eg, vaccination date) and unstructured data (eg, symptom text). The VAERS case reports should be distinguished from any other type of medical documentation (eg, discharge summaries), since both experts (physicians) and non-experts (patients and relatives) act directly as the reporters of AEs. Therefore, special processing is needed to handle the frequent non-medical syntax and semantics.

Review process

Medical officers (MOs) review all serious and death VAERS reports manually. Specifically, they review the unstructured free text fields to identify the clinical entities in a given case report, decide upon the acquisition of additional information (eg, request a copy of the medical records), and consider whether any regulatory action is warranted. VAERS reports are coded with Medical Dictionary for Regulatory Activities (MedDRA) preferred terms (PTs).[19] Non-medical data-entry personnel apply PTs to terms in AE reports according to coding conventions and algorithms; the codes are not considered to be medically confirmed diagnoses. MOs may screen and select case reports based on MedDRA codes, but they cannot fully rely on them for the analysis of safety data, mainly due to the MedDRA limitations. The inability of MedDRA to automatically group PTs with similar meanings from different system organ classes makes PT based searches incomplete unless they are based on a validated standardized MedDRA query, which are resource intensive to develop.[20] The process is performed in two steps (figure 1), both of which are laborious and time-consuming. Step 1 involves manual review of case reports, which can number in the thousands, while step 2 involves review of medical records and other documentation for a smaller number of possible cases (see example for anaphylaxis; figure 1). Here, we incorporated a variety of TM techniques and automated classifiers to reliably substitute the manual classification of anaphylaxis case reports at the first step and, thus, reduce human effort.

Brighton case definitions: the example of anaphylaxis

The Brighton Collaboration (https://brightoncollaboration.org) develops standardized, widely disseminated, and globally accepted case definitions for a large number of adverse events following immunizations (AEFIs). Each case definition is developed in a strict process that is monitored by a specific international working group of up to 20 experts and, among other items, incorporates a systematic literature search and evaluation of previous findings.[21] Based on certain criteria, the Brighton Collaboration defines the patterns that should be discovered in the reports of a surveillance system. Often, MOs try to match them with the reported symptoms in each case report (or the medical record at step 2). The Brighton Collaboration has developed a case definition for anaphylaxis (see online supplementary appendix 1), which is an acute hypersensitivity reaction with multi-organ-system involvement and can rapidly progress to a life-threatening reaction.[22] Common causes for anaphylaxis include allergens, drugs, and immunizations.[23] According to the Brighton case definition for anaphylaxis, specific major and minor criteria are described per organ system; MOs try to discover these criteria in each case report, fit them into a pattern, and classify the report as anaphylaxis or not. For example, when they read the report 'immediately after vaccination the patient presented with face edema, difficulty breathing, red eyes, wheezing, and localized rash at site of injection; also complained for weakness and reported fever 2 days before vaccination' they classify it as potentially positive primarily because there are at least two organ systems involved: dermatologic (face edema, red eyes) and respiratory (difficulty breathing, wheezing). The described 'rash' is localized and should not be considered as a dermatologic criterion, while neither 'fever' or 'weakness' are related to the vaccination or included in the case definition. It should be mentioned that the above narrative would alarm MOs even if the sudden onset (stated by 'immediately after') and the apparent rapid progression of symptoms (even though not clearly stated) were missing. Often, the time dimension is missing from VAERS reports, so MOs would still pick up this report for further review and definition of the diagnostic certainty at step 2.

Figure 1 Initially medical officers use specific MedDRA preferred terms (PT) or other keywords to extract the Vaccine Adverse Event Reporting System (VAERS) case reports of interest (usually a few thousand). Manual review requires two steps: (i) review of each case report (mainly symptom and laboratory text fields) and (ii) review of the medical record for a much smaller portion of case reports. For example, in the case of anaphylaxis, which was investigated in the current study, the PT and keyword search returned 6034 case reports that were reduced to 237 after manual review; the medical records (MR) for the latter portion of VAERS reports were obtained and reviewed resulting in 100 confirmed anaphylaxis cases.

METHODS

Corpus and validation set

We selected a subset of all the case reports that were submitted to VAERS following influenza A (H1N1) 2009 monovalent vaccines,[24] covering a period from November 22, 2009 to January 31, 2010 (N_{total}=6034). This time-window corresponded to the same period that a thorough analysis of H1N1 reports was performed following the receipt of a safety signal from Canada in mid-November.[18 25 26] Twelve MOs reviewed the reports daily (the workload share was approximately equal); their task was to use the Brighton Collaboration criteria for anaphylaxis to label them as potentially positive requiring further investigation or negative. Then, in one session with all MOs participating, the potentially positive reports were selected by consensus (N_{pos}=237); the remaining reports were classified as negative (N_{neg}=5797). MOs' classification was the gold standard for our study. Subsequently, the identification number and the symptom text field were extracted, and a class label was assigned to each case report according to the gold standard ('pos' vs 'neg' label for potentially positive vs negative reports, respectively). These data were included in a text file (one text file per report); all text files were further organized in a categorized corpus under two distinct categories ('pos' vs 'neg'). Moreover, a second set was created for validation purposes following a similar process: two MOs retrospectively reviewed (February 1, 2010–April 28, 2010) the case reports for H1N1 vaccine and created a validation set (N_{valid}=1100) with the same

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

distributional properties as the original set, that is, 4% of the case reports were potentially positive for anaphylaxis ($N_{validPos}$=44).

Feature extraction

The backbone of our work has been the TM of VAERS case reports. The starting point was the 'informative feature selection,' that is, the combination of the Brighton Collaboration criteria with the MOs' experience. Three feature representations were used:

1. Important keywords (first feature representation−>keywords).
2. Major and minor criteria that included one or more of the above keywords; MOs considered 'epinephrine' to be equal to a major criterion. Adding the feature representing the diagnosis of anaphylaxis (sometimes stated in a case report) to the feature space, a set of low-level patterns were defined (second feature representation−>low-level patterns).
3. Filtering patterns that consisted of the above criteria ('pattern1,' 'pattern2', and 'pattern3'; at least two major criteria, one major and one minor criterion, and three minor criteria, respectively). Diagnosis of anaphylaxis was also considered to be a filtering pattern alone ('pattern4') and, thus, played a dual role in our work; it should be mentioned that the proportion of cases that were detected based on the explicit diagnosis of anaphylaxis were equal to 9% (16 out of 178) in the training set, 3% (2 out of 59) in the testing set and 9% (4 out of 44) in the validation set. All filtering patterns were treated as high-level patterns (third feature representation−>high-level patterns).

Text mining processes and rule-based classifier

First, the free text in our corpus was processed using the appropriate NLP methods. Second, we worked in two directions by: (i) creating the list of lemmas (hereafter called dictionary) to represent the keywords of interest (first feature representation−>keyword−>lemma), and (ii) developing the anaphylaxis lexicon, building the grammar, tagging and parsing the free text. The grammar rules supported the extraction of major and minor criteria (second feature representation) and patterns (third feature representation) from the case reports. Using these patterns, the corresponding part of the algorithm (ie, the rule-based classifier) classified the reports into potentially positive and negative. The technical details of these processes are presented in online supplementary appendix 2.

Supervised machine learning classifiers

A number of machine learning (ML) classifiers that have been previously found to be the appropriate solutions for TC problems were trained; most of them have been also used for medical TC (binary or not): naive Bayes (NB),[5] maximum entropy (ME),[27] decision trees (DT),[6] recursive partitioning classification trees (RPCT),[28] boosted trees (BT),[29] weighted support vector machines (*w*-SVM),[3 4] SVM for sparse data (*s*-SVM),[30] stochastic boosting,[31] multivariate adaptive regression splines (MARS),[32] regularized discriminant analysis (RDA),[33] random forests (RF),[34] generalized additive model (GAM),[31] and weighted k-nearest neighbors (*w*-kNN).[35] The splitting rules that were used by the decision tree classifiers (DT, RPCT, and BT) should not be confused with the advanced grammar rules that represented the filtering patterns and were used by the rule-based classifier, as described above and presented in online supplementary appendix 2. For this reason, the decision tree classifiers and the rule-based classifier could not form a homogeneous group in a hypothetical comparison with the other classifiers.

An issue with ML classifiers is that they assign equal weights to all classes, no matter their rarity; this may affect their performance in favor of the commonest class. Thus, we included a weight parameter into the training process that allowed us to handle the problem of class imbalance; for instance, the libSVM tool (library for SVM) allows the addition of the weight parameter to the training process.[36] The weight for each class was calculated by the formula proposed by Cohen[4]:

$$w_{class} = (N_{total} - N_{class})/N_{total}$$

where N_{total} is the total number of cases and N_{class} the number of cases per class. Our weighted approach was also applied to BT, GAM, and *s*-SVM.

Python (v 2.6.4), several packages in R-statistics (v 2.11.1), and the libSVM tool were used for the training of binary classifiers, as well as for the calculation of metric values in the testing and validation sets.

Evaluation metrics and statistical analysis

For evaluating the performance of class-labeling by the classifiers, we used the macro-averaging of standard recall (R→macro-R), precision (P→macro-P), and F-measure (F→macro-F); macro-R and macro-P are preferable here since they are dominated by the more rare 'pos' category.[37] We analyzed macro-P and macro-R using Friedman's test that avoids the normality assumption and analyzes the ranks of classifiers within the data sets.[38] Moreover, the test is appropriate for use with dependent observations, which is the case here, because classifiers were tested using the same data sets. The original level of significance of 0.05 was adjusted by Bonferroni correction to account for multiplicity in testing (thus, the level of the test was 0.0125, adjusted further for multiple comparisons to 0.001).

We also performed an exploratory error analysis by presenting the mean misclassification error rate (mean-MER) for each classifier over the data set. Each classifier's SE is computed first on each data set (testing or validation) using the formula:

$$\left[\frac{\widehat{p}_{ij}(1-\widehat{p}_{ij})}{n_i}\right]^{1/2}$$

where n_i, i=1,2 is 1508 or 1100, respectively, and $\widehat{p}_{ij}$, j=1,…,13 is the error rate of the classifier on the testing and validation set. The mean-MER is obtained by averaging the individual data set dependent error rates and the associated SE is computed using a formula that adjusts for the different sizes of the data sets.

Furthermore, the impact of specific features on the classification was evaluated by computing the information gain for each unique lemma, low- and high-level pattern; information gain is employed as a term-goodness criterion for category prediction and is based on the presence or absence of a term in a document.[39] The FSelector package in R-statistics (v 2.11.1) and the training set were used for the corresponding calculations.

RESULTS

Text mining results

The application of our TM techniques and the vectorization of the TM results are shown in figure 2; the same processes were applied to all reports (both in the corpus and the validation set). The bag-of-lemmas representing a case report was compared against the dictionary and a vector of binary values was constructed to indicate the presence or absence of dictionary

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

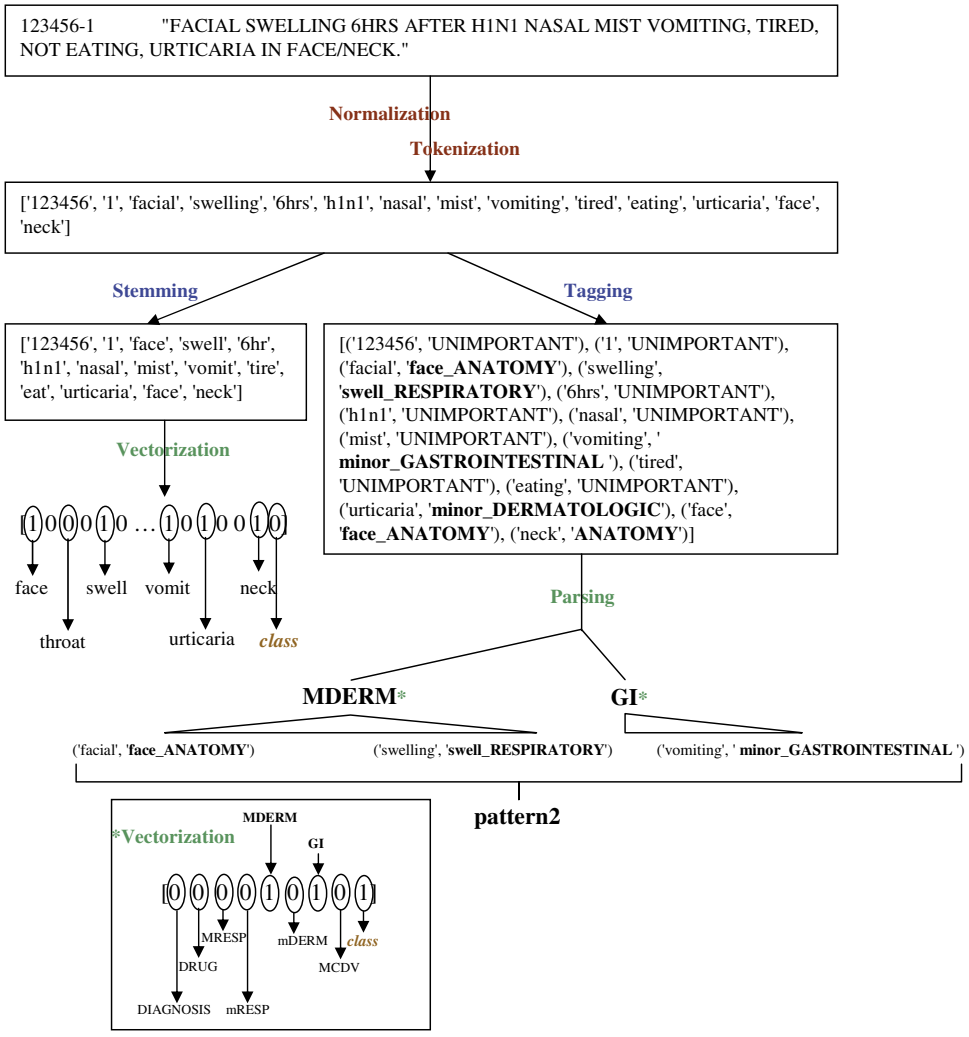

Figure 2 An example of text mining processes for a case report for anaphylaxis using either the dictionary (left branch of the diagram) or the lexicon and the grammar rules (right branch of the diagram). The output of each case report is a vector of lemmas (type I vector), a vector of low-level patterns (type II vector), or a set of high-level patterns. The two types of vectors are extended by one position to include the class label for the report. The rule-based classifier classifies this report as potentially positive based on the identification of a high-level pattern ('class'=1, ie, potentially positive). GI, gastrointestinal; MCDV, major cardiovascular; MDERM, major dermatologic; mDERM, minor dermatologic; MRESP, major respiratory; mRESP, minor respiratory.

entries in the case report; the vector (hereafter called type I vector) was extended by one position to include the class label of the report. The whole process is presented in the left branch of figure 2.

The semantic tagger assigned tags to the lemmas of each case report, while the parser interpreted the grammar to fit each tagged lemma into a rule; the rules were executed in order. Thus, the parsing process returned the low- and high-level patterns (figure 2, right branch). Again, a binary vector was created to indicate the presence or absence of the low-level patterns in each class-labeled report; each vector (hereafter called type II vector) had nine positions corresponding to the eight low-level patterns plus the class label. As for the high-level patterns, no vectorization was performed since the rule-based classifier processed this output directly by classifying a case report as potentially positive when the parser output matched any of 'pattern1,' 'pattern2', 'pattern3', or 'pattern4'; the report was otherwise classified as negative. The rule-based classification was straightforward and was applied to the case reports of all sets without getting into any training process.

On the other hand, in order to examine ML classifiers, all vectors (both type I and type II) of the original set were randomly split into a training and a testing set following the 75%–25% splitting rule (N_{train}=4526 and N_{test}=1508, respectively). Both sets had the same distributional properties, that is, 4% of the case reports were potentially positive for anaphylaxis. The performance of rule-based and ML classifiers over the testing and validation sets is presented in tables 1–3. Supplementary table 1 presents the average (over testing and validation sets) sensitivity and specificity and their associated standard errors, as well as the positive and negative predictive value of the best (across metrics) performing machine learning classifiers and rule-based classifier. Note here the equivalent performance of these classifiers in terms of all metrics presented. Yet, the rule-based classifier saved considerable computational time since it did not have to be trained, a process that was followed for *w*-SVM and BT, that is, the best ML classifiers. Moreover, the *w*-SVM and BT classifiers were also using features extracted from the Brighton Collaboration case definition and that, to a large extent, accounts for their good performance. Further analysis was performed for macro-averages and MER (see next paragraph).

Quantitative and qualitative error analysis

The null hypothesis of no difference, on average, among the ML classifiers in terms of macro-R for lemmas was rejected by Friedman's test (test statistic=12.37; p value=0.000 058, with $F_{12,12}$ distribution). Bonferroni corrected multiple comparisons (a=0.0125) indicated that BT and *w*-SVM classifiers performed best in terms of macro-R but were statistically equivalent to NB, RDA, or GAM. Notice that macro-R is equivalent to a function of sensitivity. MER analysis indicated that the mean-MER associated with BT and *w*-SVM classifiers was 0.08 (SE: ±0.0053) and 0.1015 (SE: ±0.006), respectively; this was 2.5 and 3.2 times

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

Table 1 Macro-averaged metrics for ML classifiers' performance on the first feature representation (lemmas); the ranks of classifiers are included in parentheses

	Testing set			Validation set		
Classifiers	**macro-R**	**macro-P**	**macro-F***	**macro-R**	**macro-P**	**macro-F***
NB	0.794 (10)	0.753 (4)	0.773	0.671 (11)	0.697 (4)	0.684
ME	0.701 (6)	0.863 (9)	0.773	0.577 (7.5)	0.775 (10.5)	0.661
DT	0.609 (2)	0.891 (12)	0.724	0.544 (1.5)	0.767 (9)	0.637
RPCT	0.642 (4)	0.872 (10)	0.740	0.544 (1.5)	0.732 (6.5)	0.624
BT	0.881 (13)	0.639 (2)	0.741	0.789 (12)	0.648 (2)	0.711
w-SVM	0.871 (12)	0.619 (1)	0.724	0.809 (13)	0.619 (1)	0.701
s-SVM	0.642 (4)	0.855 (7.5)	0.734	0.555 (4)	0.795 (12)	0.654
SB	0.708 (8)	0.836 (6)	0.767	0.574 (5)	0.677 (3)	0.622
MARS	0.707 (7)	0.809 (5)	0.755	0.576 (6)	0.733 (8)	0.645
RDA	0.831 (11)	0.649 (3)	0.729	0.640 (10)	0.702 (5)	0.669
RF	0.642 (4)	0.855 (7.5)	0.734	0.589 (9)	0.817 (13)	0.684
GAM	0.751 (9)	0.885 (11)	0.813	0.577 (7.5)	0.775 (10.5)	0.661
w-kNN	0.576 (1)	0.892 (13)	0.700	0.554 (3)	0.732 (6.5)	0.631

*Friedman's test indicated no statistically significant differences between the classifiers.
BT, boosted trees; DT, decision trees; F, F-measure; GAM, generalized additive model; MARS, multivariate adaptive regression splines; ME, maximum entropy; ML, machine learning; NB, naive Bayes; P, precision;
R, recall; RPCT, recursive partitioning classification trees; SB, stochastic boosting; *s*-SVM, SVM for sparse data; RDA, regularized discriminant analysis; RF, random forests; *w*-kNN, weighted k-nearest neighbors; *w*-SVM, weighted support vector machines.

higher than the smallest mean-MER among the equivalent classifiers (table 4). Additionally, the higher variability of these two classifiers as compared with the remaining classifiers should be noted. Therefore, higher macro-R comes at the expense of a higher mean-MER (and associated SE). Similarly, the null hypothesis of no differences, on average, among the ML classifiers in terms of macro-P was rejected by Friedman's test (test statistic=4.2346; p value=0.0092). Bonferroni adjusted multiple comparisons, indicated that the worst performance, in terms of macro-P, was achieved by BT and *w*-SVM classifiers, while GAM exhibited the best performance. GAM, RF, *w*-KNN, ME, DT, RPCT, and *s*-SVM were statistically equivalent in terms of macro-P.

Similar results were obtained for low-level patterns. Specifically, BT and *w*-SVM classifiers performed best in terms of macro-R with NB, ME, RDA, and GAM being statistically equivalent to BT and *w*-SVM. Our error analysis indicated higher mean-MER for the latter classifiers (for BT and *w*-SVM these were 0.0815 (±0.0055) and 0.1015 (±0.0059), respectively). These rates were two and three times higher than the minimum mean-MER obtained from the group of equivalent to the best performing classifiers. In terms of macro-P, BT and *w*-SVM exhibited the worst performance, while best performers were RF and MARS. Statistically equivalent to the latter were GAM, *w*-KNN, DT, ME, SB, *s*-SVM, and RPCT. The error analysis attributed a mean-MER of 0.032 (SE: ±0.0035) to RF and MARS, which is the smallest among all.

For the rule-based classifiers, the mean-MER over the testing and validation set was 0.0585 (SE: ±0.00465). This error rate was smaller than the best performing ML classifiers, while the performance in terms of macro-recall/precision was equivalent. Moreover, the mean macro-R was 0.8690 (SE: ±0.00 674), the mean macro-P was 0.6875 (SE: ±0.00 919), and the mean macro-F was 0.7675 (SE: ±0.00 839).

In addition to the aforementioned quantitative error analysis, we evaluated the classification performance of the ML classifiers

Table 2 Macro-averaged metrics for ML classifiers' performance on the second feature representation (low-level patterns); the ranks of classifiers are included in parentheses

	Testing set			Validation set		
Classifiers	**macro-R**	**macro-P**	**macro-F***	**macro-R**	**macro-P**	**macro-F***
NB	0.789 (11)	0.799 (4)	0.794	0.584 (8)	0.665 (3)	0.622
ME	0.707 (7.5)	0.809 (5.5)	0.755	0.587 (10)	0.719 (6)	0.646
DT	0.667 (3)	0.871 (7)	0.756	0.576 (3.5)	0.716 (5)	0.638
RPCT	0.651 (1)	0.877 (8)	0.747	0.555 (1)	0.795 (11)	0.654
BT	0.869 (12)	0.672 (2)	0.758	0.894 (13)	0.635 (2)	0.743
w-SVM	0.883 (13)	0.628 (1)	0.734	0.882 (12)	0.629 (1)	0.734
s-SVM	0.710 (9)	0.905 (11)	0.796	0.577 (5.5)	0.752 (8)	0.653
SB	0.693 (5.5)	0.899 (9.5)	0.783	0.576 (3.5)	0.733 (7)	0.645
MARS	0.677 (4)	0.944 (12)	0.789	0.567 (2)	0.816 (12)	0.669
RDA	0.788 (10)	0.783 (3)	0.786	0.585 (9)	0.683 (4)	0.630
RF	0.661 (2)	0.962 (13)	0.783	0.578 (7)	0.833 (13)	0.683
GAM	0.707 (7.5)	0.809 (5.5)	0.755	0.589 (11)	0.791 (10)	0.675
w-kNN	0.693 (5.5)	0.899 (9.5)	0.783	0.577 (5.5)	0.775 (9)	0.661

*Friedman's test indicated no statistically significant differences between the classifiers.
BT, boosted trees; DT, decision trees; F, F-measure; GAM, generalized additive model; MARS, multivariate adaptive regression splines; ME, maximum entropy; ML, machine learning; NB, naive Bayes; P, precision; R, recall; RPCT, recursive partitioning classification trees; SB, stochastic boosting; *s*-SVM, SVM for sparse data; RDA, regularized discriminant analysis; RF, random forests; *w*-kNN, weighted k-nearest neighbors; *w*-SVM, weighted support vector machines.

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

Table 3 Macro-averaged metrics for rule-based classifier's performance (high-level patterns); MER was also calculated

Data set	macro-R	macro-P	macro-F	MER
Testing set	0.889	0.691	0.777	0.058
Validation set	0.849	0.684	0.758	0.059

F, F-measure; MER, misclassification error rate; P, precision; R, recall.

using the area under the ROC curve (AUC). There are several compelling reasons as to why the AUC is appropriate for quantifying the predictive ability of classifiers, one reason being the use of AUC for testing whether predictions are unrelated to true outcomes is equivalent to using the Wilcoxon test. The best performing classifiers when AUC was used were BT and *w*-SVM. The AUC of BT when lemmas were used in the testing and validation set was, respectively, 0.9234 (95% CI 0.9198 to 0.9270) and 0.7888 (95% CI 0.7771 to 0.8006). The corresponding values of the AUC and 95% CI when low-level patterns were used for the BT classifier and for the testing and validation sets, respectively, were 0.8686 (95% CI 0.8636 to 0.8737) and 0.8944 (95% CI 0.8896 to 0.8992). For the *w*-SVM and for lemmas in the testing and validation sets we obtained AUC of 0.8706 (95% CI 0.8662 to 0.8750) and 0.8087 (95% CI 0.7989 to 0.8185). These values indicate equivalent performance of BT and *w*-SVM classifiers in terms of their predictive ability to identify truly positive reports.

We also performed a qualitative error analysis aimed at understanding if any patterns are present that allow the misclassification of reports as negative when they are truly positive. The total number of false negative reports, in both the testing and validation sets, returned by the three best performing classifiers (rule-based, BT, and *w*-SVM) was 36 (16 and 20 in the testing and validation sets, respectively). None of these reports contained the word 'anaphylaxis' and all were missing an equivalent diagnosis. MOs examined, at a second stage, these 36 reports and reclassified 15 of them as truly positive (2 and 13 in the testing and validation sets, respectively). Recall that the total number of truly positive reports in both sets equals 103. Our best performing rule-based classifier falsely misclassified only seven of these reports as negative. The symptom text of the 15 reports and more details about their qualitative error analysis are included in online supplementary appendix 3.

Table 4 Mean misclassification error rate (mean-MER) over the testing and validation data sets and the associated SE for the different ML classifiers in the case of lemmas and low-level patterns

	Lemmas		Low-level patterns	
Classifiers	Mean-MER	SE	Mean-MER	SE
NB	0.0410	0.0039	0.0380	0.0038
ME	0.0335	0.0036	0.0365	0.0037
DT	0.0355	0.0037	0.0355	0.0037
RPCT	0.0355	0.0037	0.0345	0.0036
BT	0.0800	0.0053	0.0815	0.0055
w-SVM	0.1015	0.0060	0.1015	0.0059
s-SVM	0.0350	0.0036	0.0325	0.0035
SB	0.0370	0.0038	0.0335	0.0036
MARS	0.0360	0.0037	0.0320	0.0035
RDA	0.0590	0.0045	0.0380	0.0038
RF	0.0340	0.0036	0.0315	0.0035
GAM	0.0315	0.0035	0.0345	0.0036
w-kNN	0.0370	0.0037	0.0325	0.0035

BT, boosted trees; DT, decision trees; F, F-measure; GAM, generalized additive model; MARS, multivariate adaptive regression splines; ME, maximum entropy; ML, machine learning; NB, naive Bayes; P, precision; R, recall; RPCT, recursive partitioning classification trees; SB, stochastic boosting; *s*-SVM, SVM for sparse data; RDA, regularized discriminant analysis; RF, random forests; *w*-kNN, weighted k-nearest neighbors; *w*-SVM, weighted support vector machines.

The feature analysis identified six lemmas ('epinephrin,' 'swell,' 'tight,' 'throat,' 'anaphylaxi,' 'sob') with very similar information gain results. The diagnosis of anaphylaxis and epinephrine were among the most predictive features too, when they were treated as low-level patterns, along with major and minor respiratory and major dermatologic criteria. Regarding high-level patterns all ('pattern1,' 'pattern2', and 'pattern4' that represented the diagnosis of anaphylaxis) were equivalent in terms of information gain, excluding 'pattern3'. The latter should be attributed to the extremely low (1.12%) or zero frequency of 'pattern3' in the potentially positive and negative reports of the training set, respectively. In all the other cases, the frequency of the most important predictive features was remarkably higher in the subset of the potentially positive reports.

DISCUSSION

In this study, we examined the effectiveness of combining certain TM techniques with domain expert knowledge in the case of VAERS for TC purposes; to our knowledge, no previous efforts have been reported for TM and medical TC in VAERS or any other SRS, despite the fact that various methods have been applied before to other data sources showing the potential for AE identification. For example, NLP methods have been applied to discharge hospital summaries[17] and other data mining methods to structured EHR data.[40–42] Our validated results showed that TM in any level could effectively support TC in VAERS. For example, rule-based, BT, and *w*-SVM classifiers appeared to perform well in terms of macro-R, still with some MER cost. A simple calculation over 10 000 reports for two classifiers (eg, *w*-SVM and NB for low-level patterns) with MERs of 0.10 and 0.04, respectively, would show an actual difference of 600 misclassified (either as potentially positive or as negative) reports between them. The actual cost in terms of extra workload would be those misclassified as potentially positive (based on our data that would be equal to 494 reports), but the actual cost in terms of safety surveillance would be those misclassified as negative (ie, 106 reports). Based on our error analysis, <7% (7 out of 103) of the true positive cases would be falsely classified as negative for our best performing classifier, that is, the rule-based classifier. We believe that this level of misclassification, in the context of the extensive known limitations of SRS, is probably acceptable, although we hope to engage in future efforts to refine our algorithm to reduce this even more. This further illustrates that one of the important properties of a classifier that is used to identify rare adverse events is high sensitivity because it returns a smaller number of falsely negative reports.

It could be argued that our approach lacks the automated feature extraction aspect, which has been previously reported as a strategy for TC.[43] The issue of automatically extracting features that characterize the AE accurately requires care. The problem we are called to solve is the identification of rare or very rare events from the data at hand. Because features need to be related not only statistically but also causally to the outcome, informative feature selection better serves our purposes. The basis for our claim has been the availability of solid standards (ie, Brighton case definitions) that are being used by physicians in their daily practice. Accordingly, the extraction of three feature representations supported the application of our multi-level approach. Thus, we treated the case reports not only

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

as bag-of-lemmas looking for lemmas[7] (the bag-of-words approach (stemmed or unstemmed) is rather limited[13]) but also as sources of patterns (low- and high-level); we extracted these patterns to examine their role in TC.

Informative feature selection mandated not only the use of Brighton criteria but also MOs' contributions since: (i) certain criteria were not met in the case reports and should be excluded from the feature space, for example, 'capillary refill time,' (ii) non-medical words were often used by patients to describe a symptom and should be included, for example, the word 'funny' within variations of the phrase 'my throat felt funny' to describe 'itchy throat' or 'throat closure,' and (iii) other words raised a concern for further investigation even though they were not listed in Brighton definitions, for example, 'epinephrine' or 'anaphylaxis.'

Regarding our TM methods, the construction of a controlled dictionary and lexicon is considered laborious, demanding, and costly because it relies on the recruitment of human experts.[44] However, the informative development of a flexible and relatively small controlled dictionary/lexicon appeared to be very effective in our study. The same applied to the use of the dedicated semantic tagger. A part-of-speech tagger would assign non-informative tags to a span of text (ie, symptom text in VAERS) that follows no common syntax; it would not support the grammar rules either. The grammar was also built in the same context: to better serve the extraction of the feature representations and facilitate both the rule-based classification and the training of ML classifiers.

Rule-based TC systems have been criticized for the lack of generalizability of their rules, a problem defined as 'knowledge acquisition bottleneck.'[44] However, their value in handling specific conditions should not be ignored, such as in the Obesity NLP Challenge, where the top 10 solutions were rule-based.[27] ML methods are not as transparent as the rule-based systems but have been used extensively for TC.[44] Our results showed that the rule-based classifier performed slightly better, probably due to the informative feature selection. Either rule-based or ML techniques could be applied to SRS databases and allow better use of human resources by reducing MOs' workload.

It could be argued that ensembles or a cascade of classifiers or even a modified feature space would handle the classification errors. Nevertheless, the principles of our study and the nature of VAERS would require the consideration of certain aspects prior to the application of such strategies. First, the construction of the feature space was based on the domain expert contribution; short of fully automated feature extraction, any alterations (use of new lemmas or introduction of new rules) based on this feature analysis would benefit from consultation with clinical experts to increase the chance that any such modifications would lead to meaningful results. Second, a classification error will be always introduced by the MOs who decide to acquire more information for a 'suspicious' report even if it does not meet all the criteria.

The methodology that was described in this paper and the discussion of the related aspects raise the interesting question of generalizability, that is, the transfer of components to the identification of other AEFIs. The development of a broader medical lexicon and a set of basic rules could be suggested to support the extraction of all symptoms related to the main AEFIs, such as Guillain–Barre syndrome and acute disseminated encephalomyelitis. Based on these key components, other advanced rules representing the specific criteria per AEFI (as stated in the Brighton definitions and described by MOs) could classify each report accordingly.

Our study lies partly in the field of text filtering since it investigated ways to automate the classification of streams of reports submitted in an asynchronous way.[45] Generally, MOs' intention is twofold: first, to identify the potentially positive and block the negative reports (step 1); second, to further classify those that proved to be positive into more specific categories (step 2), for example, anaphylaxis case reports into levels of diagnostic certainty. This process is similar to the classification of incoming emails as 'spam' or 'non-spam' and the subsequent categorization of the 'non-spam' emails.[46–48] Here, any further categorization would require information gathering through the review of medical records that are provided in portable document format (pdf) only. The inherent difficulties related to the production of these files limit their usability and the possibility of utilizing this source remains to be investigated.

CONCLUSION

Our study demonstrated that it is possible to apply TM strategies on VAERS for TC purposes based on informative feature selection; rule-based and certain ML classifiers performed well in this context. We plan to extend our current work and to apply the same TM strategy regarding other AEs by incorporating experts' input in a semi-automated feature extraction framework.

Funding This project was supported in part by the appointment of Taxiarchis Botsis to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. We thank all the staff members who participated on the VAERS 2009-H1N1 Influenza Response Team for assisting with the acquisition, review, and analysis of the adverse event report data.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Sinha A,** Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. *J Am Med Inform Assoc* 2009;**16**:759–67.
2. **Singleton JA,** Lloyd JC, Mootrey GT, *et al*. An overview of the vaccine adverse event reporting system (VAERS) as a surveillance system. *Vaccine* 1999;**17**:2908–17.
3. **Ambert KH,** Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *J Am Med Inform Assoc* 2009;**16**:590–5.
4. **Cohen AM.** Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *J Am Med Inform Assoc* 2008;**15**:32–5.
5. **Conway M,** Doan S, Kawazoe A, *et al*. Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inform* 2009;**78**:e47–58.
6. **Farkas R,** Szarvas G, Hegeds I, *et al*. Semi-automated construction of decision rules to predict morbidities from clinical texts. *J Am Med Inform Assoc* 2009;**16**:601–5.
7. **Mishra NK,** Cummo DM, Arnzen JJ, *et al*. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. *J Am Med Inform Assoc* 2009;**16**:576–9.
8. **Ong MS,** Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *Qual Saf Health Care* 2010;**19**:e55.
9. **Savova GK,** Ogren PV, Duffy PH, *et al*. Mayo Clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc* 2008;**15**:25–8.
10. **Solt I,** Tikk D, Gal V, *et al*. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *J Am Med Inform Assoc* 2009;**16**:580–4.
11. **DeShazo JP,** Turner AM. An interactive and user-centered computer system to predict physician's disease judgments in discharge summaries. *J Biomed Inform* 2010;**43**:218–23.
12. **Yang H,** Spasic I, Keane JA, *et al*. A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc* 2009;**16**:596–600.
13. **Cohen AM,** Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**:57–71.
14. **Hazlehurst B,** Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine* 2009;**27**:2077–83.
15. **Melton GB,** Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;**12**:448–57.

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024

16. **Murff HJ,** Forster AJ, Peterson JF, *et al*. Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc* 2003;**10**:339—50.
17. **Wang X,** Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;**16**:328—37.
18. **Varricchio F,** Iskander J, Destefano F, *et al*. Understanding vaccine safety information from the vaccine adverse event reporting system. *Pediatr Infect Dis J* 2004;**23**:287—94.
19. **Brown EG.** Using MedDRA: implications for risk management. *Drug Saf* 2004;**27**:591—602.
20. **Bousquet C,** Lagier G, Lillo-Le Louet A, *et al*. Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions. *Drug Saf* 2005;**28**:19—34.
21. **Bonhoeffer J,** Kohl K, Chen R, *et al*. The Brighton Collaboration: addressing the need for standardized case definitions of adverse events following immunization (AEFI). *Vaccine* 2002;**21**:298—302.
22. **Ruggeberg JU,** Gold MS, Bayas JM, *et al*. Anaphylaxis: case definition and guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine* 2007;**25**:5675—84.
23. **Ewan PW.** ABC of allergies: anaphylaxis. *BMJ* 1998;**316**:1442.
24. **Vellozzi C,** Broder KR, Haber P, *et al*. Adverse events following influenza A (H1N1) 2009 monovalent vaccines reported to the vaccine adverse events reporting system, United States, October 1, 2009—January 31, 2010. *Vaccine* 2010;**28**:7248—55.
25. *Quality Investigation of Combo Lot Number A80CA007A of Arepanrix™ H1N1 (AS03-Adjuvanted H1N1 Pandemic Influenza Vaccine)*. Canada: Canadian Ministry of Health, 2010.
26. **Reblin T.** *AREPANRIX™ H1N1 Vaccine Authorization for Sale and Post-Market Activities*. Ontario, Canada: Canadian Ministry of Health, 2009.
27. **Uzuner O.** Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561—70.
28. **Lewis DD,** Ringuette M. A comparison of two learning algorithms for text categorization. *Third Annual Symposium on Document Analysis and Information Retrieval* 1994;**33**:81—93.
29. **Carreras X,** Marquez L. Boosting trees for anti-spam email filtering. *4th International Conference on Recent Advances in Natural Language Processing*, 2001.
30. **Platt J.** *Sequential Minimal Optimization: a Fast Algorithm for Training Support Vector Machines*. Redmond, WA: Microsoft Research, 1998. Report No.: MST-TR-98—14.
31. **Hastie T,** Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd edn. New York, NY: Springer, 2009.
32. **Stone CJ,** Hansen MH, Kooperberg C, *et al*. Polynomial splines and their tensor products in extended linear modeling. *Ann Stat* 1997;**25**:1371—425.
33. **Friedman JH.** Regularized discriminant analysis. *J Am Stat Assoc* 1989;**84**:165—75.
34. **Rios G,** Zha H. Exploring support vector machines and random forests for spam detection. *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, 2004.
35. **Han EH,** Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classification. *Advances in Knowledge Discovery and Data Mining*, 2001:53—65.
36. **Chang CC,** Lin CJ. *LIBSVM: a Library for Support Vector Machines*. Taipei, Taiwan: Department of Computer Science; National Taiwan University, 2001.
37. **Yang Y,** Liu X. *A re-examination of Text Categorization Methods*. New York, NY: ACM, 1999:42—9.
38. **Friedman M.** The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 1937;**32**:675—701.
39. **Yang Y,** Pedersen JO. *A Comparative Study on Feature Selection in Text Categorization*. University Park, PA: Citeseer, 1997:412—20.
40. **Hinrichsen VL,** Kruskal B, O'Brien MA, *et al*. Using electronic medical records to enhance detection and reporting of vaccine adverse events. *J Am Med Inform Assoc* 2007;**14**:731—5.
41. **Jha AK,** Laguette J, Seger A, *et al*. Can surveillance systems identify and avert adverse drug events? A prospective evaluation of a commercial application. *J Am Med Inform Assoc* 2008;**15**:647—53.
42. **Linder JA,** Haas JS, Iyer A, *et al*. Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. *Pharmacoepidemiol Drug Saf* 2010;**19**:1211—15.
43. **Forman G.** An extensive empirical study of feature selection metrics for text classification. *J Mach Learn Res* 2003;**3**:1289—305.
44. **Sebastiani F.** Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 2002;**34**:1—47.
45. **Belkin NJ,** Croft WB. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM* 1992;**35**:29—38.
46. **Androutsopoulos I,** Koutsias J, Chandrinos KV, *et al*. *An Experimental Comparison of Naive Bayesian and Keyword-based Anti-spam Filtering With Personal E-mail Messages*. New York, NY: ACM, 2000:160—7.
47. **Bekkerman R,** McCallum A, Huang G. Automatic categorization of email into folders: benchmark experiments on Enron and SRI corpora. *Center for Intelligent Information Retrieval, Technical Report IR*, 2004:418.
48. **Drucker H,** Wu D, Vapnik VN. Support vector machines for spam categorization. *IEEE Trans Neural Netw* 1999;**10**:1048—54.

Downloaded from https://academic.oup.com/jamia/article/18/5/631/835446 by guest on 24 April 2024