Drug Saf (2018) 41:579–590
https://doi.org/10.1007/s40264-018-0641-7

ORIGINAL RESEARCH ARTICLE

Sorting Through the Safety Data Haystack: Using Machine Learning to Identify Individual Case Safety Reports in Social-Digital Media

Shaun Comfort[1] · Sujan Perera[2] · Zoe Hudson[1] · Darren Dorrell[1] · Shawman Meireis[1] · Meenakshi Nagarajan[2] · Cartic Ramakrishnan[2] · Jennifer Fine[1]

Published online: 14 February 2018

Abstract

Introduction There is increasing interest in social digital media (SDM) as a data source for pharmacovigilance activities; however, SDM is considered a low information content data source for safety data. Given that pharmacovigilance itself operates in a high-noise, lower-validity environment without objective 'gold standards' beyond process definitions, the introduction of large volumes of SDM into the pharmacovigilance workflow has the potential to exacerbate issues with limited manual resources to perform adverse event identification and processing. Recent advances in medical informatics have resulted in methods for developing programs which can assist human experts in the detection of valid individual case safety reports (ICSRs) within SDM.

Objective In this study, we developed rule-based and machine learning (ML) models for classifying ICSRs from SDM and compared their performance with that of human pharmacovigilance experts.

Methods We used a random sampling from a collection of 311,189 SDM posts that mentioned Roche products and brands in combination with common medical and scientific terms sourced from Twitter, Tumblr, Facebook, and a spectrum of news media blogs to develop and evaluate three iterations of an automated ICSR classifier. The ICSR classifier models consisted of sub-components to annotate the relevant ICSR elements and a component to make the final decision on the validity of the ICSR. Agreement with human pharmacovigilance experts was chosen as the preferred performance metric and was evaluated by calculating the Gwet AC1 statistic (gKappa). The best performing model was tested against the Roche global pharmacovigilance expert using a blind dataset and put through a time test of the full 311,189-post dataset.

Results During this effort, the initial strict rule-based approach to ICSR classification resulted in a model with an accuracy of 65% and a gKappa of 46%. Adding an ML-based adverse event annotator improved the accuracy to 74% and gKappa to 60%. This was further improved by the addition of an additional ML ICSR detector. On a blind test set of 2500 posts, the final model demonstrated a gKappa of 78% and an accuracy of 83%. In the time test, it took the final model 48 h to complete a task that would have taken an estimated 44,000 h for human experts to perform.

Conclusion The results of this study indicate that an effective and scalable solution to the challenge of ICSR detection in SDM includes a workflow using an automated ML classifier to identify likely ICSRs for further human SME review.

Shaun Comfort and Sujan Perera contributed equally to the work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40264-018-0641-7) contains supplementary material, which is available to authorized users.

✉ Shaun Comfort
comforts@gene.com

[1] Genentech, A Member of the Roche Group, Roche, South San Francisco, CA, USA

[2] IBM Watson Health, Cambridge, MA, USA

Key Points

A machine learning classifier achieved substantial agreement with a human expert when classifying social digital media posts as valid individual case safety reports

This level of performance could not be achieved with a conventional rule and dictionary approach to classification

Combining a machine learning approach with human review has the potential to be an effective and scalable solution to the challenge of identifying individual case safety reports within social digital media posts

1 Introduction

Safety surveillance in the premarket clinical trial process is designed to identify common adverse events (AEs) and drug reactions (ADRs) occurring in study populations. Typical clinical development programs include sample sizes between a few hundred to several thousand study patients in total; allowing for identification of AEs occurring between approximately 3% (i.e., 3/100) down to 0.3% (i.e., 3/1000) by the 'Rule of Three' [1]. However, many subsequent ADRs are identified after the drug is on the market due to factors including exposure to an expanded patient population, concomitant medication use, dosing patterns, off-label usage, and intentional misuse [2, 3]. Effective postmarket pharmacovigilance, defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" [4], relies on swift, accurate, and comprehensive reporting of ADRs through the submission of individual case safety reports (ICSRs) to the appropriate regulatory bodies. A streamlined, global approach to pharmacovigilance increases the power of signal detection activities to identify and supplement the initial safety profile based on randomized controlled trials in the area of clinical safety. This is facilitated by the application of the revised Guideline for Clinical Safety Data Management: Data Elements for Transmission of ICSRs (E2B), which was developed by the International Council for Harmonization and has been widely adopted as the standard for ICSR reporting [5].

At a minimum, a valid ICSR must contain at least one of each of the following four elements: (i) an identifiable patient, (ii) an identifiable reporter, (iii) a suspected drug, and (iv) an AE [6]. The currently accepted practice is for pharmaceutical companies to review all spontaneous reports submitted directly to the company or any of its employees [7]. This may also include AEs detected during post-market studies, reported in the scientific literature, or identified during evaluation of data that the company collected for unrelated purposes. This improves the depth and breadth of data available for signal detection compared with randomized controlled trials, but there remain well known limitations including under-reporting and selective reporting [8, 9].

In recent years, social-digital media (SDM) posts have been proposed as a potential new source of data for ADRs which can compensate for some of the limitations in the existing spontaneous ICSR reporting system [10, 11]. Individuals commonly turn to the internet to find peer communities through which they can gain insight into their condition and with whom they can share their experiences with therapeutics [12, 13]. These social media exchanges and posts have the potential to offer a wellspring of insight into potential risks and benefits of therapeutics and patient education needs. The complexity, richness, diversity, and scale of this content presents not only an opportunity for garnering new insights into patient perspectives but also a challenge to the current capacity for prompt reading and digestion of this information [14].

Previous publications have shown that the majority of digital media posts do not meet the minimum requirements of a reportable ICSR, and the transient and anonymous nature of the medium make follow up on incomplete posts challenging, if not impossible [15]. In spite of this, the large pool of SDM posts still has the potential to enable the industry to detect or strengthen other safety signals. The nature of social media offers a unique, direct, unaltered reflection of a patient's experience and is more likely to reveal safety-relevant information in their own words (e.g., ''drug X makes me feel angry''), than adverse reactions diagnosed in a clinical trial setting. There is some evidence that useful insights into drug misuse and abuse can also be garnered from SDM, which may otherwise go undetected in the more traditional post-market safety data [16]. For these and other reasons, it is worth pursuing evaluations of technology with the capability to pick out valuable safety information from vast quantities of data.

Current EU regulations require marketing authorization holders to track and report only ADRs derived from SDM posts within their own website forums or otherwise brought to their attention [16–18]. While the current regulatory landscape does not mandate the pharmaceutical industry to review noncompany-controlled SDM (as in 'active surveillance'), the regulator's expectations to monitor such data sources may change if a viable solution to the problem of scaling is found.

Computer automation is one potential mechanism for creating a scalable solution for monitoring SDM sources in pharmacovigilance activities; however, traditional rule-based algorithms perform poorly at tasks that require parsing natural language [19]. Machine learning (ML) is an alternative programming approach that is widely used for natural language processing and sentiment analysis. ML has been evaluated for early detection of safety signals from social media sources and in labeling ADRs within Twitter posts [15, 20–22]. In this study, we investigated the potential of ML to identify valid ICSRs relative to human subject matter expert (SME) assessments of the same SDM data.

Development of a program that can identify potential AEs and determine if they are valid ICSRs within the highly noisy platform of social media requires a deep understanding of both pharmacovigilance and ML algorithms. To address this complexity, we established a collaboration between Roche and IBM with the goal of determining if ML could be leveraged to screen SDM posts for potential ICSRs and reduce the burden on human pharmacovigilance experts without sacrificing accuracy.

2 Methods

2.1 Scope

The objective of this proof-of-concept effort was to develop and evaluate a computer program which could classify English-language SDM posts as valid versus invalid ICSRs. The objective was to start with an entirely rule- and dictionary-based approach to classification and strategically add in ML elements to address identified deficiencies in performance. Due to the regulatory impact of false-negative results and its intended role as a pre-screening tool, we prioritized minimization of type II errors (false negatives) at the expense of increased type I errors (false positives) during the development phase. Our approach was to optimize for high agreement between the automated classifier and a human SME.

Valid ICSRs were those that contained, within the text of the post, (i) an identifiable patient, (ii) an identifiable reporter, (iii) a suspected drug, and (iv) an AE. Content available through embedded links, pictures or non-text sources will not be used to determine ICSR validity.

For this proof-of-concept study, we applied the following limits to the project scope. First, the final classifier will only assess posts it identifies as English language. Second, it will only identify pure ICSRs (i.e., those containing AEs related to noxious or unintended effects, off-label use, overdose, misuse, abuse or medication errors). AEs related to lack of effect or disease progression were considered out of scope due to the complexity and subjectivity required for determining reportability.

2.2 Data Collection and Management

A separate team within Roche had previously collected a dataset of 311,189 SDM posts using the social media brand monitoring platform Radian6 to identify posts that mention Roche products and brands in combination with common medical and scientific terms [23]. The social media outlets from which these posts were collected included Twitter, Tumblr, Facebook, and a spectrum of news media blogs. Search terms included key words associated with the pharmaceutical industry, Roche brand, and senior personnel. The lists included Roche product names, pharmaceutical terminology (e.g., diabetes, oncology, drug approval, FDA, influenza) and brands specific to Roche (e.g., Chugai, Genentech). Negative searches were also applied such 'NOT Roche-Posay' (a separate cosmetic company). The sourcing of data from the internet was indiscriminate towards language and resulted in a dataset consisting of over 44 different languages; however, the majority (55–60%) were in English. A full list of search terms can be found in electronic supplementary material 1.

A member of the Roche team (DD) loaded the full dataset into an Oracle database and programmatically searched and labeled the posts for Roche product names and MedDRA preferred terms (PT) and lowest level terms (LLT). This full dataset was supplied to both the IBM team, to facilitate understanding of the dataset, and to the Roche pharmacovigilance team for further processing and curation. For training and testing purposes, DD randomly selected three non-overlapping subsets (Set A, B, and C) from the source data (Fig. 1). For each set, we have identified the number of posts excluded because either the human pharmacovigilance SME or the software identified them as non-English. The remaining posts were identified as either valid or invalid ICSRs according to the methods described in Sect. 2.3.

2.3 Ground Truth and Discrepancy Analysis

We established ground truth for set A by having three pharmacovigilance SMEs (ZH, SC, and SM) independently review the posts to determine if they met the criteria of an ICSR. The SMEs reviewed all posts in the development set regardless of whether a Roche product name or MedDRA LLT or PT had been programmatically identified. We then labeled posts as valid ICSRs if at least two out of three SMEs flagged it as a complete ICSR.

For sets B and C, we established ground truth by having the posts evaluated by a single pharmacovigilance SME (ZH), who is also the Roche global expert for SDM. This

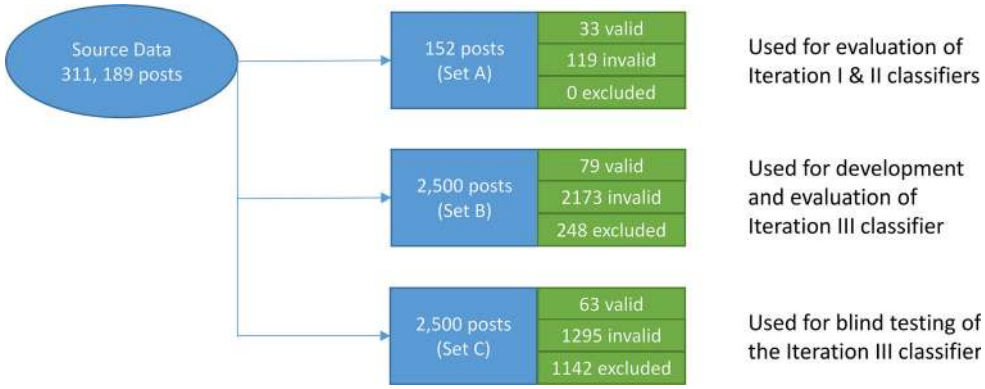


Fig. 1 Breakdown of source data and curated subsets. Blue boxes indicate the data batches received by the software development team in various stages of development. The green boxes contain the split between valid, invalid, and excluded individual case safety reports (ICSRs) in the respective dataset. Posts were excluded because they fell outside the scope of the proof-of-concept study (see Sect. 2)

reflected how an ML system would ideally be used in the pharmacovigilance workflow, to supplant the need for an initial human review of all social media posts. Since only a single operator was used to establish ground truth for the final test data, we had the two other pharmacovigilance SMEs (SC and SM) and one ML SME (SP) perform a subsequent discrepancy analysis to verify the accuracy of the human operator and, where possible, to categorize the reasons for both false positives and false negatives by the classifier.

2.4 Structure of Individual Case Safety Report (ICSR) Classifiers

All solutions presented in the paper utilize the same high-level approach to structuring an ICSR classifier (Fig. 2). First, three annotators identify the presence of the minimum required ICSR entities: an adverse event, a drug, and a patient. In social media, the reporter is assumed to be the author of the post. The fourth annotator identifies relationships between the identified entities, and then the ICSR detector takes the original post and annotated entities as inputs and outputs the ICSR decision for the post. We iteratively developed the components of the ICSR classification framework. This is due to three reasons: (i) availability of the SME annotated datasets, (ii) to establish and communicate a common understanding between Roche and the IBM team about the nuances of the problem, and (iii) to receive feedback from pharmacovigilance SMEs for the ICSR classification results.

2.5 Iteration I: Rule-Based ICSR Classifier

The Iteration I model used a set of dictionaries and a rule-based approach to identifying potential ICSRs. Each of the four annotators used a simple text matching approach to identify AEs, drugs, patients, and relationships. The AE annotator in this model used the MedDRA dictionary and

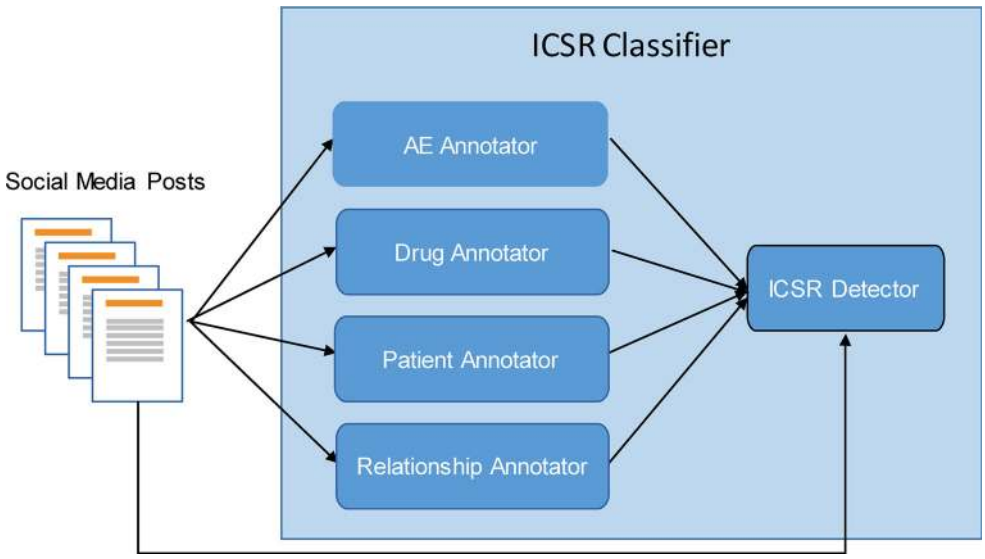


Fig. 2 Components of the individual case safety report (ICSR) classification framework. *AE* adverse event

terms like *cause, deteriorated, worsened* and *aggravate* to identify the AEs. The drug annotator dictionary included all generic and brand names for Roche pharmaceutical products, and the patient annotator included variations of 80 pronouns such as I, my, mine, you, adult, patient, baby, boy, and girl. The relationship annotator applied simple rules to determine likely relationships between the three elements above. For example, a text pattern like *'DRUG_NAME cause MedDRA_TERM'* would suggest that a medical condition described by MedDRA_TERM is an AE reported to be caused by DRUG_NAME.

In this first approach, the ICSR detection module was also developed as a rule-based solution. The following are some of the rules that were applied to detect ICSRs in iteration I:

If all ICSR elements are not present:
  Not ICSR
If all ICSR elements are present:
  Patient mention is not a first person pronoun:
    Low confidence ICSR
  Patient mention is first person pronoun but has a weak relationship with other entities:
    Low confidence ICSR
  ICSR elements show strong relationship:
    High confidence ICSR

Weak and strong relationships were characterized by the absence or presence, respectively, of an explicit relationship between the entities. If the entities were present within the same sentence, but our pattern-based algorithm did not find an explicit relationship, then they were marked as having a weak relationship; whereas entities within the same sentence with explicit relationships were marked as having strong relationships.

2.6 Iteration II: Machine Learning (ML) Approach to Adverse Event (AE) Annotation

In Iteration II, we supplemented the rule-based AE annotator with a machine-learned AE annotator. All other modules from iteration I remained unchanged. To train the new AE annotator, we used a publically available Twitter dataset of 1784 tweets previously annotated for adverse events [21]. We selected an independent dataset for training the AE annotator to prevent overfitting of the ML model. We specifically selected an annotated Twitter dataset as it is closest to the grammatical, morphological, and syntactic properties of the SDM posts that we are focusing on in this study, and ML models are very sensitive to the linguistic features exhibited by the text on which they are trained [24]. We trained an instance of KnIT pipeline [20–22] to detect adverse events in tweets by exploiting their syntactic and semantic features. A more in-depth explanation of the ML methods can be found in electronic supplementary material 2. Example AEs from the annotated training corpus are shown below.

- Drug A destroyed my entire body
- Drug A nearly killed me
- Drug A made me hungry, dizzy, and tired
- Drug A knocked me out

2.7 Iteration III: ML Approach to ICSR Detector

For Iteration III, we upgraded the ICSR detector from a rule-based approach to an ML model. All other modules were kept the same from iteration II to iteration III. We used a support vector machine (SVM) algorithm to develop the new ICSR classifier (electronic supplementary material 2). Before training the classifier, we processed the input text to identify the portion of the text that potentially contained the information related to the ICSR decision. Namely, we extracted the sentences or text snippets of the digital media posts that contained any drug or adverse event mention as identified by the annotators. We call such snippets 'focus text'. Then we extracted commonly used syntactic and semantic features from the focus text to train the SVM. The features we used include *n*-grams, dictionary look-up based features, word clustering based on word vector embeddings, and brown clusters [25–28]. For training and stability testing of the iteration III classifier, we combined data subsets A and B to create a curated set of 2404 posts. Of these, 112 were valid ICSRs, and 2292 were invalid ICSRs.

From this set, we built five cross-validation sets by randomly assigning 170 posts to a training set, with a fixed distribution of 80 valid ICSRs and 90 invalid ICSRs, and assigning the remaining 2234 posts to the testing set. Robustness of the classifier was evaluated by performing a training and testing iteration on each of the five cross-validation sets [29].

Before locking the Iteration III classifier for the final performance test on subset C, it was trained on the entire validated, combined data from subsets A and B.

2.8 Testing Method and Performance Metrics

Each version of the classifier was tested by predicting the ICSR results to a subset of the available dataset and comparing the model's results to the ground truth established by the SMEs. An early challenge we encountered was establishing a common language for describing performance because the domains of ML and pharmacovigilance use different terms to refer to the same underlying statistics. For instance, the typical table for displaying agreement between two assessment methods is referred to as a confusion matrix in the ML community but is known

Table 1 Performance metrics

Name	Value
True positive (tp)	No. of true positives[a]
True negative (tn)	No. of true negatives[a]
False positive (fp)	No. of false positives[a]
False negative (fn)	No. of false negatives[a]
Accuracy (Acc)	$\frac{tp+tn}{tp+fp+tn+fn}$
Gwet AC1	$\frac{Acc-e(\pi)}{1-e(\pi)}$
$e(\pi)$	$\left\{\frac{(2tp+fp+fn)/2}{tp+fp+tn+fn}\right\}^2+\left\{\frac{(2tn+fp+fn)/2}{tp+fp+tn+fn}\right\}^2$
Area under the curve	Trapezoidal method [30]

SME subject matter expert

[a]True positive, negative, etc are based on SME-determined 'ground truth'

as a 2 × 2 contingency table in the pharmacovigilance community. To ensure clarity of communication between both domains, we built a conversion table for the two sets of statistics and agreed upon three metrics for evaluating an ICSR classifier: accuracy, area under the receiver operator characteristic curve (AUC), and Gwet AC1 (Table 1 and electronic supplementary material 3).

Accuracy is equivalent to the overall percent agreement and is a useful metric when evaluating the performance of classifiers with a rule-based ICSR detector. In contrast, for the iteration III ML classifier, which provides a probability value for each decision, AUC is a more appropriate metric for evaluating performance [30]. For assessing agreement between ground truth annotations and predictions generated by our solution, we selected the Gwet AC1 statistic instead of the more common Cohen's kappa. Cohen's kappa is a robust metric when the number of positive and negative test elements are roughly equivalent. As the ratio skews away from 1:1, the kappa statistic becomes highly sensitive to single mismatches. The Gwet AC1 statistic is a reliable alternative statistic for measuring agreement that is less sensitive to the ratio of positive and negative test elements [31, 32].

2.9 Fermi Estimation Human ICSR Identification Time

As a final exercise, we undertook a simple 'Fermi' analysis to estimate the likely time it could take human SMEs to manually evaluate the entire bulk of SDM posts in this case study [33]. Our approach consisted of combining an estimate of the range of word volume per digital media posts (estimated max and min of 10–10,000 words/post) with the international human reading speed of 184 ± 29 (SD) words/min [34]. We used the geometric mean (e.g., the square root of the product of upper and lower bounds) approach to 'Fermi Problems' to estimate the mode. The Program Evaluation and Review Technique (PERT) was then used to model the likely human ICSR identification time in minutes and hours, for the cumulative dataset, as an approximate β function.

3 Results

3.1 Assessment of the Dataset

The data collection method collected 311,189 SDM posts between 2012 and 2016. Roughly 80.47% were posted in 2016, 19.50% were posted in 2015, and 0.03% were from before 2015. Because vernacular on social media varies between sites, posts were aggregated from a variety of sources to ensure the classification models were appropriately challenged (Table 2).

A manual review revealed four common reasons that the majority of the posts would not be classified as ICSRs, despite containing both an adverse event term and a drug name. First, the mentioned AE term may be the primary condition being treated by the drug rather than a separate adverse event occurring as a consequence of the drug. Second, the post may be describing a positive patient experience. Third, it may be an advertisement that happens to include AE terms unrelated to a drug name. Fourth, the post may be a general discussion of side effects but not an individual experience. A small selection of contrived valid and invalid SDM ICSRs can be found in Table 3.

3.2 Iterative Development of an ICSR Classifier

We built the Iteration I ICSR classifier by defining rules over the presence of four required ICSR entities and their relationships. When tested on a small dataset of 152 posts, set A, the Iteration I classifier achieved an accuracy of 65% and a Gwet AC1 of 46% (Fig. 3a). Critically, it missed nearly two-thirds of the valid ICSRs according to the ground truth established by our subject matter experts (Fig. 3b).

Table 2 Breakdown of the social media sources of the core dataset

Social media site	Number of posts
Twitter	168,745
Online news and blogs	106,336
Tumblr	32,961
Facebook	2754
YouTube	142
Other	251
Total	311,189

Table 3 Contrived examples of social media posts containing valid and invalid ICSRs

Valid ICSRs	Invalid ICSRs
Got the *DRUG NAME* **burn** going on today	This **flu** has laid me out, ready to get back to life…Thanks *DRUG NAME*
I **feel like death** after taking *DRUG NAME*	*ROCHE* **Cancer** *drug* success adds to pressure on competitor
I took *DRUG NAME* and now I'm about to **pass out**	The most common side effects of *DRUG NAME* are **nausea** and **vomiting**

The terms in italics indicate drug keywords and the terms in bold indicate potential adverse event keywords and phrases

ICSR individual case safety report

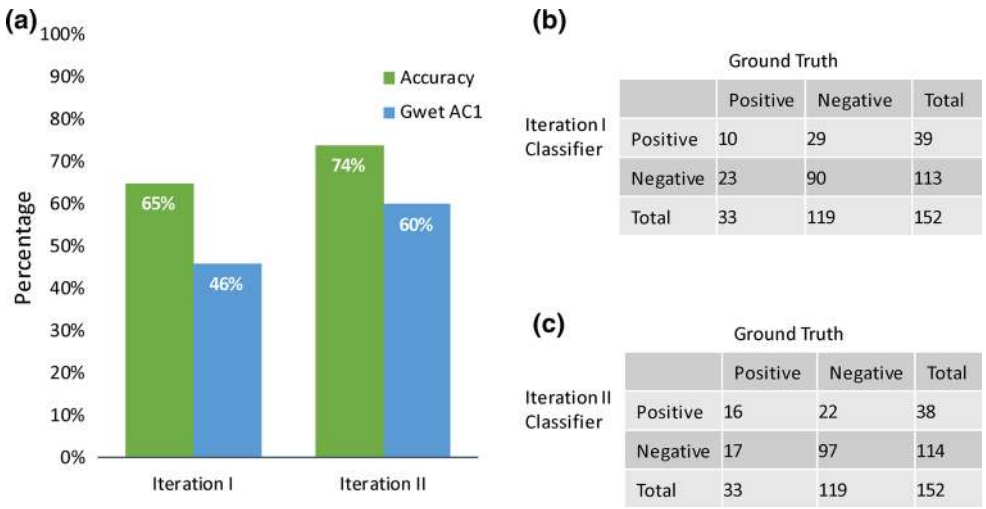


Fig. 3 Performance of Iteration I and Iteration II classifiers. **a** Graph of accuracy and Gwet AC1 for both classifiers. **b** Confusion matrix for Iteration I classifier. **c** Confusion matrix for Iteration II classifier

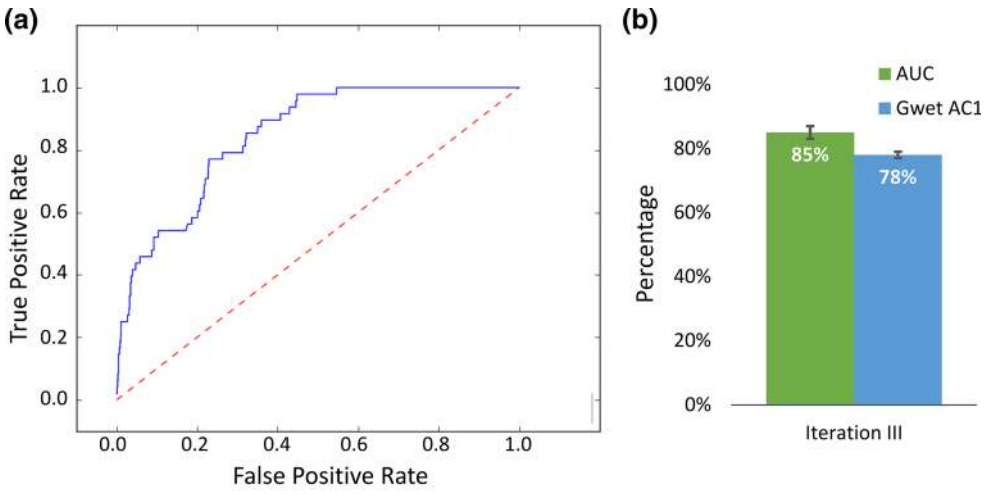
An evaluation of the missed ICSRs showed that the dictionary-based approach to annotating AEs was particularly ineffective. Specifically, the annotator failed to recognize or misinterpreted colloquialisms and slang; therefore, for the Iteration II classifier, we adopted an ML approach to annotating AEs. With the addition of the ML AE annotator, the Iteration II classifier achieved an accuracy of 74% and a Gwet AC1 of 60% (Fig. 3a). The confusion matrix shows a reduction in both false positives and false negatives (Fig. 3c); however, the performance is only moderate according to the standard benchmarking of agreement [35]. Common performance metrics have been calculated for each confusion matrix and are reported in electronic supplementary material 4.

The earlier assessment of the dataset revealed that the presence of all four elements was only marginally associated with identification of a valid ICSR. Analysis of the performance of the iteration II classifier indicated that the rule-based ICSR detector was ineffective at making the types of distinctions identified in Sect. 3.1**. Therefore, the next step in development was to replace the rule-based ICSR detector with an ML solution. The new ICSR classifier could produce the probability values for its classifications. This allowed us to plot the receiver operator characteristic curve (Fig. 4a) and calculate the AUC as a measure of accuracy. We present the results averaged over fivefolds of the dataset C. This classifier achieved an AUC of $85 \pm 2\%$ and a Gwet AC1 of $78 \pm 1\%$ (Fig. 4b). The full confusion matrix of the average performance on the five cross-validation sets is reported in Fig. 4c.

3.3 Testing the ML Classifier

Next, we evaluated the iteration III model's classification of the 2500 posts that were not part of the training set (set C—blind dataset). This is a critical component of validating an ML model, which ensures that it can perform equally well on new data and is not overfit to the training set. Of the 2500 posts, we excluded 1139 because either the automatic language detection tool or the human reviewer identified them as non-English. An additional three posts were removed because they fell outside the scope of our experiment because the ICSR type was a lack of effect or disease progression (see Sect. 2.1).

The agreement between the iteration III classifier and the human social media pharmacovigilance SME, as measured by Gwet AC1, for the remaining 1358 posts was 78%. The full confusion matrix of results is reported in Fig. 5a and common performance metrics have been



(c)

Iteration III Classifier	Ground Truth: Positive	Negative	Total
Positive	20 ± 1.9	328 ± 26.7	348 ± 28.6
Negative	12 ± 1.9	1873 ± 26.7	1885 ± 28.6
Total	32	2201	2234

Fig. 4 Performance of Iteration III classifier. **a** Plot of the receiver operator characteristic (ROC) curve of the Iteration III classifier. **b** Graph of area under the ROC curve and Gwet AC1 for the Iteration III classifier (average ± SD). **c** Confusion matrix of the five cross-validation results for the Iteration III classifier (average ± SD). *AUC* area under the curve, *SD* standard deviation

calculated and are reported in electronic supplementary material 4.

We reviewed the 234 discrepant results and grouped the reason for discrepancy into eight buckets (Fig. 5b). The top three reasons, which accounted for over two-thirds of false-positive results, were unreported rationale for AE by the classifier, AE term was contained within a word salad, and term reported by the classifier as an AE was not an AE. Unreported rationale refers to all the posts where the ML ICSR detector incorrectly calculated a high probability that the post contained an AE even though the AE annotator did not identify an AE. Of the five valid ICSRs missed by the iteration III classifier, two were comparatively long (392 and 1403 words), diluting the significance of individual features, one had no explicit mention of a patient, one was missed by the rules-based drug annotator, and one used an unclear antecedent to link the patient to the drug.

As a final test, we ran the full dataset of 311,189 unannotated posts through the ML ICSR classifier. In its current design, the algorithm was able to exclude over 88% of the posts in < 48 h. In contrast, our Fermi/PERT model for plausible human curation time of the same dataset predicts an expected effort of approximately 44,000 work hours or 22 full-time individuals working for 2000 h over 1 year (Table 4). While the uncertainty range of our estimate is large, the result suggests that human identification of ICSRs from large volumes of digital material can require significant resources.

4 Discussion

This proof-of-concept study demonstrated both the utility and comparability of an ML approach to human SMEs for screening SDM posts for potential ICSRs. The initial iteration model established that a strict rule- and dictionary-based approach to ICSR classification resulted in an accuracy of only 65%. Of particular concern was the fact that the initial rules-based model missed over two-thirds of the valid ICSRs. Subsequently adding an ML-based AE annotator for the second iteration improved the accuracy and agreement and this performance was further improved with the addition of an ML ICSR detector. On the final blind test set of 2500 posts, the Gwet AC1 agreement between the social media pharmacovigilance SME and the ML classifier was 78%. The Iteration III model was successful at identifying 92% of the valid ICSRs from a highly diverse and noisy social media dataset. One noted observation is that the ICSR classifier was able to correctly identify the ICSRs even when the annotators fell short in identifying respective entities (false negatives) or incorrectly identified entities (false positives). This proves that errors in upstream processes have minimal impact on the overall classification task. This is due to the fact that the ICSR detector has the capability to analyze the linguistic features of the SDM post and determine its validity/invalidity. The output of the annotators are just a few factors among many in deriving the final outcome.

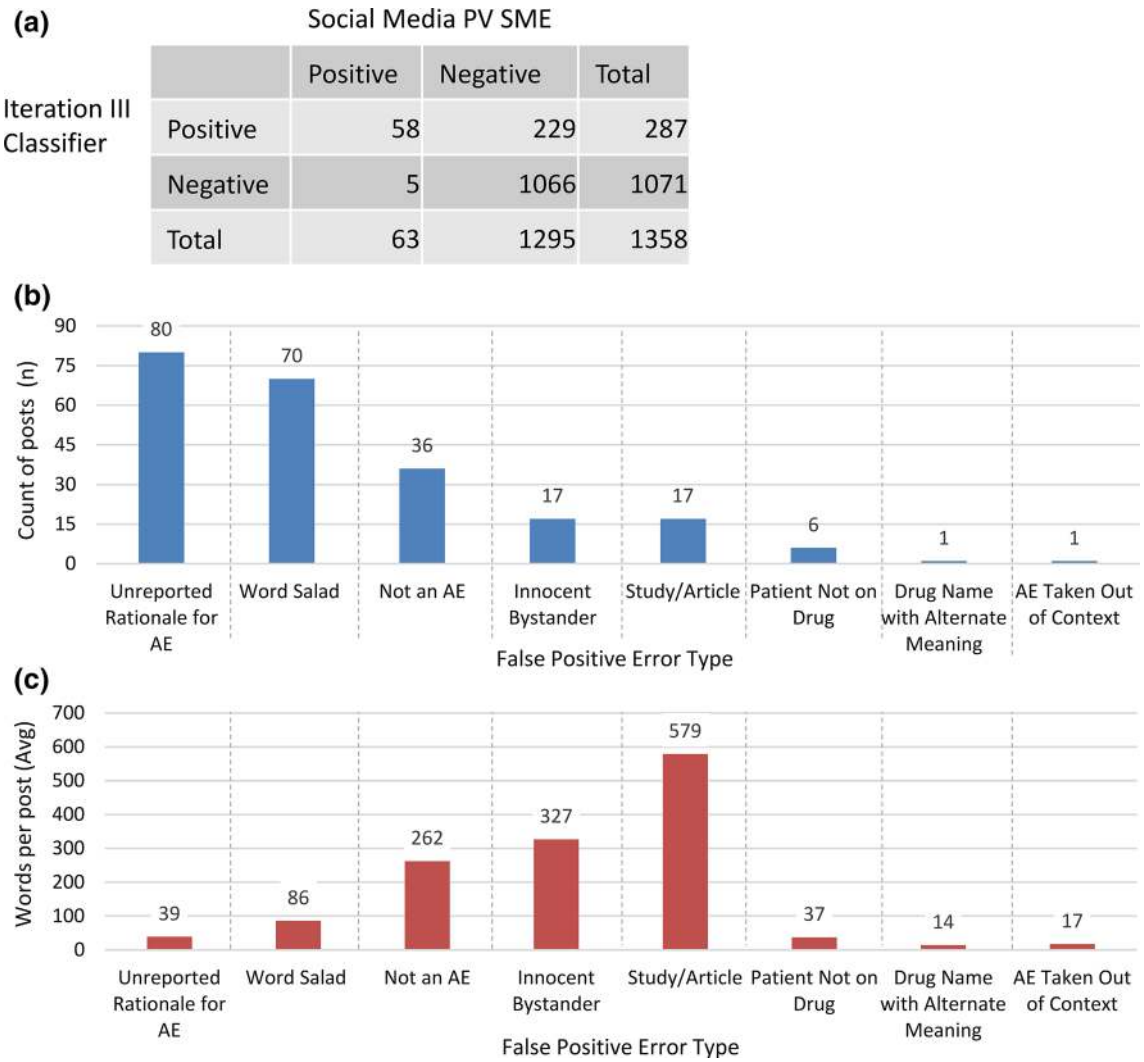


Fig. 5 Performance of the Iteration III classifier on a blind set. **a** Confusion matrix of the blind testing set results for the Iteration III classifier. **b** Chart of count of false-positive results by reason. **c** Chart of average length of post by reason for false-positive result. *AE* adverse event, *Avg* average

Table 4 Program evaluation and review technique estimate of time for a human to evaluate the digital media data collection

Variables	Minimum	Maximum	Mode	Exp	SD	L90%CI*	U90%CI*
Posts (*n*)	311,189	311,189	311,189	311,189	0.0	311,189	311,189
Human reading speed (wpm)	136	232	178	180	16	153	206
Post length (words)	10	10,000	316	1879	1665	48	2101
Read/ID speed (min/post)	0.07	43.10	1.78	8.38	7.17	0.31	10.22
Total evaluation time (min)	22,882	13,413,319	554,001	2,608,701	2231,701	96,517	3,179,934
Total evaluation time (h)	381	223,555	9233	43,478	37,196	1609	52,999

Exp exponential, *ICSR* individual case safety report, *L90%CI* lower 90% confidence interval, *SD* standard deviation, *U90%CI* upper 90% confidence interval, *wpm* words per minute

* PERT Beta Function Approximate Confidence Intervals

Social digital media presents many challenges for ICSR identification for both manual and automated ICSR curation. Human curation is the current gold standard for identifying valid ICSRs; however, the speed at which SDM posts are generated, the bulk of which are often advertisements, nonsense (e.g., 'word salad'), or unrelated,

makes a manual human-only process tedious and impractical for curation of large datasets. It may be that the introduction of specific mobile applications for AE reporting, such as the recent successful public–private Web-RADR partnership, will make capturing safety events from SDM users much easier [36]. However, it is unclear what proportion of the total volume of SDM reporting would go into these well designed reporting vehicles. Thus, approaches that can sift through large volumes of raw material to identify likely ICSRs are still useful.

Classical computer algorithms can rapidly process high volumes of data; however, the nuances of tone, style, slang, and context make non-ML-based algorithms insufficient for uniquely identifying valid ICSRs. We believe that the results of this effort indicate that an effective and scalable solution includes a workflow using an automated classifier to detect likely ICSRs for automatic classification or further human SME review.

Finally, during evaluation of the final Iteration III model, we identified several opportunities for further model development, likely resulting in a performance improvement. Specifically, the inclusion of additional filters for the removal of nonsense posts and enhanced language identification would address recurrent issues in parsing scrambled or irrelevant SDM posts from the more likely valid ICSR posts. It may also be valuable to train the AE annotator on more data from sources outside of Twitter, which accounted for a large proportion of this project's dataset. The 140-character length restriction of Twitter posts leads to altered linguistic patterns which may negatively impact the annotator's ability to identify AEs in other less restricted forums [37, 38]. Another challenge to overcome is the impact of heavy social media users, which could lead to biased signal detection. Within a single forum, it may be possible to prevent duplicate reporting by tracking the commenter's user IDs, but across platforms or in venues that allow anonymous handles, other strategies must be developed and employed to prevent 'over sampling' of one or a few individual's SDM posts.

Today we are looking at SDM as a growing source of potential insight into patients' experiences with illness and therapeutics. Efforts such as the Innovative Medicines Initiative WEB-RADR project, conducted by a consortium of health authorities (e.g., MHRA, EMA, etc), industry (e.g., Novartis, Bayer, etc), and academia, are beginning to develop a clear regulatory framework of how SDM could be used to advance patient safety [36]. This is compounded by a 2016 study which suggested that both general public (83%) and caregiver participants (63%) felt that SDM was feasible for patient safety reporting [39]. However, the complexity, richness, diversity, and volume of this data requires novel technologies to digest it quickly, accurately, and at scale.

While this pilot focused on safety-related insights, other insights into benefits, adherence, educational needs, preferences, and patient behaviors could enhance our ability to not only detect safety issues quickly but also to support patients more appropriately and completely and to even detect new benefits. These social media outlets are likely to change, as will what we term SDM. In the future, safety data may be further supplemented by data from mobile patient safety reporting applications, wearable devices such as watches, exercise apps, and smart devices. As biometric data begins to be uploaded by users, the volume and type of data to review will grow nonlinearly. Future technical solutions will need to read unstructured text from a host of sources and combine those insights with ones from exogenous structured data sources like wearable devices, electronic medical records, claims data, and social media to present a holistic view of patients' experience with their conditions, care, therapeutics, and lifestyles to inform not only drug safety but all aspects of drug discovery and development.

4.1 Limitations

There are several limitations to this study which bear consideration. Foremost, although we had 311,189 posts at our disposal, we were limited to manual review, which meant that we were only able to establish the ground truth for a subset of 5152 posts. In addition, the frequency of valid ICSRs ranged from 1 to 5%, which is very limiting for both development and testing purposes. To truly establish performance metrics, we would need to establish the ground truth for a much larger fraction of our available dataset. Second, our dataset pulls exclusively from posts that mention Roche products or keywords and as a result are weighted towards the fields of oncology, and immunologic and metabolic disorders. Performance on a broader dataset cannot be easily extrapolated from performance on a Roche-exclusive dataset. Specifically, the ICSR detector module in the Iteration III classifier would need to be retrained on a more diverse dataset before being applied to wider use. Third, the initial filtering software incorrectly identified several posts as non-English, which had to be excluded from analysis. Depending on the accuracy of the language-filtering software, it is possible that we excluded relevant posts and introduced selection bias. Performance may be improved with further optimization of language filtering.

5 Conclusions

In this paper, we presented the development of an ML tool that we envision could potentially be used to prescreen SDM posts for potential valid ICSRs. The final classifier

had a sensitivity of 92.1% and a specificity of 82.3%. Additional work must be done to improve the positive predictive value to further reduce the burden on a human workforce without increasing the false-negative rate. This may be achieved through the application of gated filters based on post length or other characteristic features and training of the AE annotator on sources outside of Twitter.

Author Contributions SC, SP, ZH, DD, CR and MN contributed to the study design. All authors contributed to data collection, curation and/or analysis. SC, SP, ZH, SM, and JF drafted the manuscript. All authors approved the final manuscript.

Acknowledgements Medical writing support was provided by Jessamine Winer-Jones, PhD, IBM Watson Health. Content review was provided by Elenee Argentinis, JD, IBM Watson Health. Technical Advice, IT, and PM support was provided by Krishna Pendurthi and Mukund Chauhan from Roche Safety Informatics as well as from John Unterholzner, Sheng Hua Bao, PhD, Jacob Schutz, PhD, Van C Willis, PhD and Albert Lee from IBM.

Funding Funding for development and testing of the classifiers was supplied by Roche.

Compliance with ethical standards

Conflict of interest Shaun Comfort, Zoe Hudson, Darren Dorrell, Shawman Meireis, and Jennifer Fine were employed by Roche at the time this research was completed. Sujan Perera, Cartic Ramakrishnan, and Meena Nagarajan were employed by IBM at the time this research was completed.

Ethics statement All human subject data used in this analysis were publicly available and used in a de-identified format whenever possible.

References

1. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249:1743–5.
2. Härmark L, van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. Eur J Clin Pharmacol. 2008;64:743–52.
3. Brewer T, Colditz GA. Postmarketing surveillance and adverse drug reactions: current perspectives and future needs. JAMA. 1999;281:824–9.
4. World Health Organization. The importance of pharmacovigilance. 2002. http://apps.who.int/medicinedocs/en/d/Js4893e/. Accessed 15 Sep 2017.
5. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. E2B(R3) electronic transmission of individual case safety reports implementation guide—data elements and message specification. 2014. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/guidances/ucm274966.htm. Accessed 15 Sep 2017.
6. Cobert B. Cobert's manual of drug safety and pharmacovigilance. Sudbury: Jones & Bartlett Publishers; 2011.
7. Food and Drug Administration. 21 CFR 314.80: postmarketing reporting of adverse drug experiences. 2017. https://www.ecfr.gov/cgi-bin/text-idx?SID=db68ad73ff4f35bdb5750e78aebfd5b5&mc=true&node=se21.5.314_180&rgn=div8. Accessed 15 Sep 2017.
8. Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Determinants of under-reporting of adverse drug reactions. Drug Saf. 2009;32:19–31.
9. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions. Drug Saf. 2006;29:385–96.
10. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inf. 2015;54:202–12.
11. Duh MS, Cremieux P, Audenrode MV, Vekeman F, Karner P, Zhang H, et al. Can social media data lead to earlier detection of drug-related adverse events? Pharmacoepidemiol Drug Saf. 2016;25:1425–33.
12. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res. 2010;12:e19.
13. Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen Intern Med. 2011;26:287–92.
14. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res. 2013;15:e85.
15. Bhattacharya M, Snyder S, Malin M, Truffa MM, Marinic S, Engelmann R, et al. Using social media data in routine pharmacovigilance: a pilot study to identify safety signals and patient perspectives. Pharm Med. 2017;31:167–74.
16. Anderson LS, Bell HG, Gilbert M, Davidson JE, Winter C, Barratt MJ, et al. Using social listening data to monitor misuse and nonmedical use of bupropion: a content analysis. JMIR Public Health Surveill. 2017;3:e6.
17. European Medicines Agency. Guideline on good pharmacovigilance practices: module VI—management and reporting of adverse reactions to medicinal products. 2012. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf. Accessed 15 Sep 2017.
18. Lengsavath M, Pra AD, Ferran A-Md, Brosch S, Härmark L, Newbould V, et al. Social media monitoring and adverse drug reaction reporting in pharmacovigilance. Therap Innov Regul Sci. 2017;51:125–31.
19. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge: MIT Press; 1999.
20. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inf Assoc. 2017;24:813–21.
21. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inf Assoc. 2015;22:671–81.
22. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, et al. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf. 2017;40:317–31.

23. Salesforce Marketing Cloud Radian6. 2016. https://www.marketingcloud.com/products/social-media-marketing/radian6/. Accessed 15 Sep 2017.
24. Derczynski L, Maynard D, Rizzo G, van Erp M, Gorrell G, Troncy R, et al. Analysis of named entity recognition and linking for tweets. Inf Process Manag. 2015;51:32–49.
25. Baker LD, McCallum AK. Distributional clustering of words for text classification. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval; 1998, vol. 21, p. 96–103.
26. Gamon M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on computational linguistics; 2004, vol. 20, p. e841.
27. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:1301.3781 [cs.CL].
28. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics. 2007;23:2768–74.
29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc of the 14th international conference on artificial intelligence; 1995, vol. 14, p. 1137–45.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
31. Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Stat Methods Inter-rater Reliab Assessm. 2002;1:1–6.
32. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61:29–48.
33. Weinstein L, Adam JA. Guesstimation: solving the world's problems on the back of a cocktail napkin. Princeton: Princeton University Press; 2009.
34. Trauzettel-Klosinski S, Dietz K. Standardized assessment of reading performance: the new international reading speed texts IReST standardized assessment of reading performance. Invest Ophthalmol Vis Sci. 2012;53:5452–61.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
36. European Medicines Agency Innovative Medicines Initiative WEB-RADR Workshop Report. Mobile technologies and social media as new tools in pharmacovigilance 2016. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2017/02/WC500221615.pdf. Accessed 15 Sep 2017.
37. Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, et al. Part-of-speech tagging for twitter: Annotation, features, and experiments. Proceedings of the 49th annual meeting of the association for computer linguistics: human language technology: short papers; 2011, vol. 49, p. 42–7.
38. Han B, Baldwin T. Lexical normalisation of short text messages: makn sens a #twitter. Proceedings of the 49th annual meeting of the association for computer linguistics: human language technology; 2011, vol. 1, p. 368–78.
39. Omar I, Harris E. The use of social media in ADR monitoring and reporting. J Pharmacovigil. 2016;4:1–9.