Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task

Md Sarker, Emory University
Maksim Belousov, University of Manchester
Jasper Friedrichs, Infosys Ltd
Kai Hakala, University of Turku
Svetlana Kiritchenko, National Research Council Canada
Farrokh Mehryary, University of Turku
Sifei Han, University of Kentucky
Tung Tung, University of Kentucky
Anthony Rios, University of Kentucky
Ramakanth Kavuluru, University of Kentucky

Only first 10 authors above; see publication for full author list.

Journal Title: Journal of the American Medical Informatics Association
Volume: Volume 25, Number 10
Publisher: Oxford University Press (OUP): Policy B - Oxford Open Option B | 2018-10-01, Pages 1274-1283
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/jamia/ocy114
Permanent URL: https://pid.emory.edu/ark:/25593/tvcdb

Final published version: http://dx.doi.org/10.1093/jamia/ocy114

Copyright information:
© The Author(s) 2018. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nd/4.0/).

Accessed October 15, 2019 2:47 AM EDT
Research and Applications

Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task

Abeed Sarker,1 Maksim Belousov,2 Jasper Friedrichs,3 Kai Hakala,4,5 Svetlana Kiritchenko,6 Farrokh Mehryary,4,5 Sifei Han,7 Tung Tran,7 Anthony Rios,7 Ramakanth Kavuluru,7,8 Berry de Bruijn,6 Filip Ginter,4 Debanjan Mahata,9 Saif M Mohammad,6 Goran Nenadic,2 and Graciela Gonzalez-Hernandez1

1Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, 2School of Computer Science, University of Manchester, Manchester, UK, 3Infosys Limited, Palo Alto, California, USA, 4Turku NLP Group, Department of Future Technologies, University of Turku, Turku, Finland, 5The University of Turku Graduate School, University of Turku, Turku, Finland, 6Digital Technologies Research Centre, National Research Council Canada, Ottawa, Canada, 7Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA, 8Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, Kentucky, USA, and 9Bloomberg, New York, New York, USA

Corresponding Author: Abeed Sarker, PhD, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive 421A, Philadelphia, PA 19104, USA (abeed@pennmedicine.upenn.edu)

Received 8 May 2018; Revised 18 July 2018; Editorial Decision 28 July 2018; Accepted 2 August 2018

ABSTRACT

Objective: We executed the Social Media Mining for Health (SMM4H) 2017 shared tasks to enable the community-driven development and large-scale evaluation of automatic text processing methods for the classification and normalization of health-related text from social media. An additional objective was to publicly release manually annotated data.

Materials and Methods: We organized 3 independent subtasks: automatic classification of self-reports of 1) adverse drug reactions (ADRs) and 2) medication consumption, from medication-mentioning tweets, and 3) normalization of ADR expressions. Training data consisted of 15 717 annotated tweets for (1), 10 260 for (2), and 6650 ADR phrases and identifiers for (3); and exhibited typical properties of social-media-based health-related texts. Systems were evaluated using 9961, 7513, and 2500 instances for the 3 subtasks, respectively. We evaluated performances of classes of methods and ensembles of system combinations following the shared tasks.

Results: Among 55 system runs, the best system scores for the 3 subtasks were 0.435 (ADR class F1-score) for subtask-1, 0.693 (micro-averaged F1-score over two classes) for subtask-2, and 88.5% (accuracy) for subtask-3. Ensembles of system combinations obtained best scores of 0.476, 0.702, and 88.7%, outperforming individual systems.

Discussion: Among individual systems, support vector machines and convolutional neural networks showed high performance. Performance gains achieved by ensembles of system combinations suggest that such strategies may be suitable for operational systems relying on difficult text classification tasks (eg, subtask-1).
BACKGROUND AND SIGNIFICANCE

Social media have enabled vast numbers of people, anywhere, from any demographic group, to broadcast time-stamped messages on any topic, in any language, and with little or no filter. The Pew Social Media Fact Sheet published in 2017 revealed that approximately 70% of the population in the United States actively uses social media, and the user base is seeing continuous growth globally. Their earlier research suggested that “health and medicine” is one of the most popular topics of discussion in social media, with 37% of adults identifying it as the most interesting topic. Due to the presence of vast amounts of health-related information, it is being increasingly utilized as a data source for monitoring health trends and opinions. Social media traffic is being used or considered for many health-related applications, such as public health monitoring, tracking of disease outbreaks, charting behavioral factors such as smoking, responding to mental health issues, and pharmacovigilance. The social media revolution has coincided with drastic advancements in the fields of natural language processing (NLP) and data analytics, and, within the health domain, biomedical data science. However, despite recent advances, performing complex health-related tasks from social media is not trivial. There are 2 primary hurdles along the way for such tasks: 1) picking up a signal, and 2) drawing conclusions from the signal. This paper concentrates entirely on (1), as it describes considerations and solutions for representing noisy textual messages into formalized and pure data elements. However, we briefly want to shift focus to (2). Drawing conclusions from social media signals is not without risk due to several types of bias or sources of error outside the text representations. A patient alleging an adverse drug event may be wrong (deliberately or not) on the drug intake details, the symptoms themselves (including misdiagnoses), or the attribution of causality between the drug and the alleged reaction. In addition, reporting biases may exist, varying among drugs, symptoms, or subpopulations. Despite these caveats, social media traffic is very likely to contain signals that we cannot afford to ignore. Additionally, the availability of large volumes of data makes it a rewarding resource for the development and evaluation of data-centric health-related NLP systems. While innovative approaches have been proposed, there is still substantial progress to be made in this domain. In this paper, we report the design, results, and insights obtained from the execution of a community-shared task that focused on progressing the state of the art in NLP of health-related social media text.

Shared tasks and evaluation workshops have been a popular approach for progressing NLP methods on specialized tasks. They have proven to be effective in providing clear benchmarks in rapidly evolving areas. Their benefits to participating researchers include a reduction in their individual data annotation and system evaluation overhead. The benefits to the field include the very objective evaluation, using standardized data, metrics, and protocols. Successes of general-domain NLP shared tasks, such as Computational Natural Language Learning (CONLL), Text Analysis Conference (TAC), and the International Workshop on Semantic Evaluation (SemEval) have inspired domain-specific counterparts. In the broader medical domain, these include BioASQ, BioCreative, CLEF eHealth, and i2b2, which have significantly advanced health-related NLP.

Through the Social Media Mining for Health (SMM4H) shared tasks, we aimed to further extend these efforts to NLP from health-related social media. While medical text is itself complex, text originating from social media presents additional challenges to NLP, such as typographic errors, ad hoc abbreviations, phonetic substitutions, use of colloquial language, ungrammatical structures, and the use of emoticons. For text classification, data imbalance due to noise and usage of non-standard expressions typically lead to the underperformance of systems on social media texts. Concept normalization from this resource, which is the task of assigning standard identifiers to text spans, is among the least explored topics. Within medical NLP, tools utilizing lexicons and knowledge bases such as MetaMap and cTAKES have been used for identifying and grouping distinct lexical representations of identical concepts. Such tools are effective for formal texts from sources such as medical literature, but they perform poorly when applied to social media texts.

The SMM4H-2017 shared tasks were focused on text classification and concept normalization from health-related posts. The text classification tasks involved the categorization of tweets mentioning potential adverse drug reactions (ADRs) and medication consumption. The concept normalization task required systems to map ADR expressions to standard IDs. In this paper, we expand on the SMM4H-2017 shared task overview by presenting analyses of the performances of the systems and classes of systems, additional experiments, and the insights obtained and their implications for informatics research.

MATERIALS AND METHODS

Data and annotations

We collected all the data from Twitter via the public streaming API, using generic and trade names for medications, along with their common misspellings, totaling over 250 keywords. For subtasks-1 and -2, the annotated datasets for training were made available to the public with our prior publications, while subtask-3 included previously unpublished data. Evaluation data were not made public at the time of the workshop. Following the completion of the workshop, we have made all annotations publicly available.

Subtask-1 included 25,678 tweets annotated to indicate the presence or absence of ADRs (these ADRs are as reported by the users and do not prove causality). The annotation was performed by 2 annotators with inter-annotator agreement (IAA) of $\kappa = 0.69$ (Cohen’s $kappa$) computed over 1082 overlapping tweets. Subtask-2 included 17,773 annotated tweets categorized into 3 classes—definite intake (clear evidence of personal consumption), possible intake (likely that the user consumed the medication, but the evidence is unclear), and no intake (no evidence of personal consumption). IAA was $\kappa = 0.88$ for 2 annotators, computed over
1276
cite{journal}

Subtask-1 was a binary text classification task for which systems were required to predict if a tweet mentions an alleged ADR or not. Such classification tasks are important because most of the medication-mentioning chatter on social media, including Twitter, is noise. Systems were evaluated on their ability to accurately detect tweets belonging to the ADR class using the class-specific F1-score metric, which is based on the true positive (tp), false negative (fn) and false positive (fp) counts:

\[
\text{recall} = \frac{tp}{tp + fn}; \quad \text{precision} = \frac{tp}{tp + fp};
\]

\[
F_1\text{-score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}
\]

Subtask-2 involved 3-class text classification, and systems were required to classify mentions of personal medical consumption from tweets, most of which do not explicitly express personal consumption. The evaluation focused on assessing systems’ abilities to detect the definite and possible cases of consumption, and thus relied on micro-averaged F1-score for the 2 classes. For subtask-3, given an ADR expression, systems were required to identify the mapping for the expression in the MedDRA vocabulary. The evaluation metric for this task was accuracy (ie, proportion of correctly identified MedDRA PTs in the evaluation set).

Methodologies and system descriptions

Subtasks-1 and -2: text classification
For subtasks-1 and -2, high-scoring systems frequently used support vector machines (SVMs), deep neural networks (DNNs), and classifier ensembles. We now provide further details of the methods, particularly focusing on the high-performing systems, and selected methods and features.

Approaches and features. For the traditional classifiers (eg, SVMs), high-performing systems utilized lexical features such as word and character n-grams, negations, punctuations, and word clusters along with specialized domain-specific and semantic features. NRC-Canada, the top-performing team for subtask-1, extended its existing state-of-the-art sentiment analysis and stance detection systems, and incorporated features such as: n-grams generalized over domain terms (ie, words or phrases representing medications from the RxNorm list or entries from the ADR lexicon) are replaced with $<MED>$ and $<ADR>$, respectively, pre-trained word embeddings, and word clusters obtained from one million tweets that mention medications. In addition, for subtask-2, the team’s systems utilized sentiment features—sentiment association scores obtained from existing manually and automatically created lexicons, including Hu and Liu Lexicon, Norms of Valence,


Arousal, and Dominance, labMT, and NRC Emoticon Lexicon. The UKNLP (University of Kentucky) systems used similar feature sets (eg, the ADR lexicon) along with 2 additional features: the sum of words’ pointwise mutual information (PMI) scores as a real-valued feature based on the training examples and their class membership; and handcrafted lexical pairs of drug mentions (subtask-2 only) preceded by pronouns (the count of first, second, and third personal pronouns with and without negation followed by a drug mention).

The system from the TurkuNLP (University of Turku) team for subtask-2 was based on an ensemble of convolutional neural networks (CNNs) applied on sequences of words and characters. The model also relied on pre-trained word embeddings and term frequency-inverse document frequency (TF-IDF) weighted bag-of-words representations with singular-value-decomposition-based dimensionality reduction. The InfyNLP team (Infosys Ltd), top-performers for subtask-2, employed double-stacked ensembles of shallow CNNs. Multiple candidate ensembles of 5 shallow CNNs were first trained, using random search for parameter optimization. The top k best performing ensembles, as per cross-validation on the training data, were then stacked to make predictions on the test set. The team used publicly available pre-trained word embeddings to represent words in the network, with no additional features or text representations. The primary differences between the team’s method and other CNN-based approaches were the use of shallow networks, while most other implementations were deep, as well as the use of random search to generate many candidate models for the double-stacked ensembles.

Strategies for addressing data imbalance. For subtask-1, a key challenge was data imbalance, as only approximately 10% of the tweets presented ADRs. NRC-Canada used undersampling to rebalance the class ratio from about 1: 10 to 1: 2. Other methods for dealing with data imbalance included cost-sensitive training (CSaRUS; Arizona State University) and minority oversampling (NTTMU; multiple universities, Taiwan), but without much success. For both classification tasks, most teams also incorporated classifier ensembles (eg, by combining votes from multiple classifier predictions or via model averaging) to improve performance over the smaller class(es).

Subtask-3: normalization
Methods utilized for subtask-3 consisted of a multinomial logistic regression model, 3 variants of recurrent neural networks (RNNs),
and an ensemble of the 2 types of models. The gnTeam (University of Manchester) performed lexical normalization to correct spellings and convert out-of-vocabulary words to their closest candidates before converting the phrases into dense vector representations using several publicly available sources. Following the generation of this representation, the team applied multinomial logistic regression, an RNN classifier with bidirectional gated recurrent units (GRUs), and an ensemble of the 2. For the ensemble, the final predictions were made based on the highest average value for each class derived from predicted probabilities of the base learners. The UKNLP systems employed a deep RNN model that realized a hierarchical composition in which an example phrase was segmented into N constituent words, and each word was treated as a sequence of characters. In contrast to gnTeam’s GRUs, their systems used long short-term memory (LSTM) units, and, for a variant of the system, utilized additional publicly available data for training.

### Baselines, ensembles, and system extensions

For each subtask, we implemented 3 baseline systems for comparison against the submitted systems. For subtasks-1 and -2, we implemented naïve bayes, SVMs, and random forest classifiers. We used only preprocessed (lowercased and stemmed) bag-of-words features, and, for the latter 2 classifiers, we performed basic parameter optimization via grid search. For subtask-3, our baseline systems relied on exact lexical matching: the first with MedDRA PTs, the second with LLTs, and the third with the training set annotations.

Following the execution of the shared task evaluations, we implemented multiple voting-based ensemble classifiers, using the system submissions as input. Our objective was to assess how combinations of optimized systems performed relative to individual systems, and to explore strategies by which system predictions could be combined to maximize performance. For subtask-1, we combined groups of system predictions (eg, all and top n), and used different thresholds of votes for the ADR class (eg, majority and greater than n votes) to make predictions. We performed a similar set of experiments for subtasks-2 and -3, and because they are multi-class problems, we used only majority voting for prediction.

Following the shared task evaluations, teams with the top-performing systems were invited to perform additional experiments using fully annotated training sets (in addition to those publicly available). This enabled the teams to experiment with different system settings and optimization methods, which were not possible earlier due to the time constraint imposed by the submission deadline. The test set annotations were shared with the selected teams privately for evaluation. Performances of these extended systems along with summaries of the extensions, relative to their reported methods in the shared task descriptions, are presented in the next section.

### RESULTS

#### Shared task system performances

Fifty-five system runs from 13 teams were accepted for evaluation (24 submissions from 9 teams for subtask-1; 26 from 10 for subtask-2; and 5 submissions from 2 for subtask-3). We categorized the methods employed by the individual submitted systems into 5 categories: CNN, SVM, RNN, Other, and Ensembles, where “Other” represents traditional classification approaches such as logistic regression and k-nearest neighbor, and “Ensembles” includes stacks of ensembles. Figure 3 shows the relative distributions of these categories of approaches employed by the submitted systems.

For individual systems, NRC-Canada’s SVM-based approach, which utilized engineered domain-specific features and parameter optimization via 5-fold cross-validation over part of the training set, obtained the highest ADR F1-score of 0.435. InfyNLP’s ensemble of shallow CNNs topped subtask-2 with a micro-averaged F1-score of 0.693. For subtask-3, all submitted systems showed similar performances, with an ensemble of RNN and logistic regression obtaining the best accuracy. Tables 1–3 present the performances of selected submissions for the subtasks, along with the performances of the baseline systems and post-workshop ensembles. We show only the top-performing systems for subtasks-1 and -2; full set of results and exclusion criteria for the shared task can be found in the overview.

<table>
<thead>
<tr>
<th>System/Team</th>
<th>ADR precision</th>
<th>ADR recall</th>
<th>ADR F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1: Naïve Bayes</td>
<td>0.774</td>
<td>0.098</td>
<td>0.174</td>
</tr>
<tr>
<td>Baseline 2: SVMs with RBF kernel</td>
<td>0.501</td>
<td>0.215</td>
<td>0.219</td>
</tr>
<tr>
<td>Baseline 3: Random Forest</td>
<td>0.429</td>
<td>0.066</td>
<td>0.115</td>
</tr>
<tr>
<td>NRC-Canada</td>
<td>0.392</td>
<td>0.488</td>
<td>0.435</td>
</tr>
<tr>
<td>CSaRUS-CNN</td>
<td>0.437</td>
<td>0.393</td>
<td>0.414</td>
</tr>
<tr>
<td>NorthEasternNLP (NorthEastern University)</td>
<td>0.395</td>
<td>0.431</td>
<td>0.412</td>
</tr>
<tr>
<td>UKNLP (University of Kentucky)</td>
<td>0.498</td>
<td>0.337</td>
<td>0.402</td>
</tr>
<tr>
<td>TsuiLab (University of Pittsburgh)</td>
<td>0.336</td>
<td>0.348</td>
<td>0.342</td>
</tr>
</tbody>
</table>

Figure 3. Percentage distributions for 5 categories of approaches attempted by teams for the shared tasks.

Table 1. Performance metrics for selected system submissions for subtask-1, baselines, and system ensembles. Precision, recall, and F1-score over the ADR class are shown. The top F1-score among all systems is shown in bold. Detailed discussions about the approaches can be found in the system description papers referenced.
Table 2. Performance metrics for selected system submissions for subtask-2, baselines, and system ensembles. Micro-averaged precision, recall, and F1-scores are shown for the definite intake (class 1) and possible intake (class 2) classes. The highest F1-score over the evaluation dataset is shown in bold. Detailed discussions about the approaches can be found in the system description papers referenced (when available)

<table>
<thead>
<tr>
<th>System/Team</th>
<th>Micro-averaged precision for classes 1 and 2</th>
<th>Micro-averaged recall for classes 1 and 2</th>
<th>Micro-averaged F1-score for classes 1 and 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1: Naïve Bayes</td>
<td>0.359</td>
<td>0.503</td>
<td>0.419</td>
</tr>
<tr>
<td>Baseline 2: SVMs</td>
<td>0.652</td>
<td>0.436</td>
<td>0.523</td>
</tr>
<tr>
<td>Baseline 3: Random Forest</td>
<td>0.628</td>
<td>0.487</td>
<td>0.549</td>
</tr>
<tr>
<td>InfyNLP³⁵ (Infosys Ltd)</td>
<td>0.725</td>
<td>0.664</td>
<td>0.693</td>
</tr>
<tr>
<td>UKNL³⁵ (University of Kentucky)</td>
<td>0.701</td>
<td>0.677</td>
<td>0.689</td>
</tr>
<tr>
<td>NRC-Canada³⁵</td>
<td>0.708</td>
<td>0.642</td>
<td>0.673</td>
</tr>
<tr>
<td>TJIIP (Tongji University, China)</td>
<td>0.691</td>
<td>0.641</td>
<td>0.665</td>
</tr>
<tr>
<td>TurkuNLP⁵⁵ (University of Turku)</td>
<td>0.701</td>
<td>0.630</td>
<td>0.663</td>
</tr>
<tr>
<td>CSaRUS-CNN⁵⁶ (Arizona State University)</td>
<td>0.709</td>
<td>0.604</td>
<td>0.652</td>
</tr>
<tr>
<td>NTTMU³⁵ (Multiple Universities, Taiwan)</td>
<td>0.690</td>
<td>0.554</td>
<td>0.614</td>
</tr>
<tr>
<td>Ensemble all: majority vote</td>
<td>0.736</td>
<td>0.657</td>
<td>0.694</td>
</tr>
<tr>
<td>Ensemble top 10: majority vote</td>
<td>0.726</td>
<td>0.679</td>
<td>0.702</td>
</tr>
<tr>
<td>Ensemble top 7: majority vote</td>
<td>0.724</td>
<td>0.673</td>
<td>0.697</td>
</tr>
<tr>
<td>Ensemble top 5: majority vote</td>
<td>0.723</td>
<td>0.667</td>
<td>0.694</td>
</tr>
<tr>
<td>Ensemble top submissions from top 5 teams: majority vote</td>
<td>0.727</td>
<td>0.673</td>
<td>0.699</td>
</tr>
</tbody>
</table>

Table 3. System performances for subtask-3, including baselines and ensembles. Summary approaches and accuracies over the evaluation set are presented. Best performance is shown in bold

<table>
<thead>
<tr>
<th>Team</th>
<th>Approach summary</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>Exact lexical match with MedDRA PT</td>
<td>11.6</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>Exact lexical match with MedDRA LL or PT</td>
<td>25.1</td>
</tr>
<tr>
<td>Baseline 3</td>
<td>Match with training set annotation</td>
<td>63.5</td>
</tr>
<tr>
<td>gnTeam⁶⁶ (University of Manchester)</td>
<td>Multinomial Logistic Regression</td>
<td>87.7</td>
</tr>
<tr>
<td></td>
<td>RNN with GRU</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>88.5</td>
</tr>
<tr>
<td>UKNL⁶⁷ (University of Kentucky)</td>
<td>Hierarchical RNN with LSTM</td>
<td>87.2</td>
</tr>
<tr>
<td></td>
<td>Hierarchical RNN with LSTM and external data</td>
<td>86.7</td>
</tr>
<tr>
<td>Ensemble</td>
<td>All systems</td>
<td>88.7</td>
</tr>
<tr>
<td></td>
<td>Top 3</td>
<td>88.7</td>
</tr>
</tbody>
</table>

DISCUSSION

In this section, we outline the findings of the error analyses performed on the top-performing systems, pointing out the key challenges that we have identified. We then summarize the insights obtained and the implications of these for health informatics research.

Error analysis

For subtasks-1 and -2, the most common reason for false negatives was the use of infrequent, creative expressions (eg, “I have metformin tummy today: -(!).” Low recall due to false negatives was particularly problematic for subtask-1, and systems also frequently misclassified rarely occurring ADRs. False positives were caused mostly by classifiers mistaking ADRs for related concepts such as symptoms (eg, “headache”) and beneficial effects (ie, “hair loss reversal”). Lack of context in the length-limited posts poses prob-
lems for annotators as well as the systems. For subtask-1, the relatively low IAA results from ambiguous expressions of ADRs without clear contexts (eg, “headache & xanex :(!”). The IAA results in systems having low performance ceilings for this subtask and also suggests that the annotations in the dataset may not be completely reliable, as judgments made in the absence of supporting information are often subjective. Better representations of the posts (eg, with supporting context) and future improvements in core NLP methods specialized for social media texts may result in improved performances in downstream tasks such as classification, by enabling systems to better capture contexts and dependencies. For subtask-2, additional common causes for misclassification were inexplicit mentions about medication consumption, or explicit consumption mentions without clear indications about who took the medication. Instances of the “possible intake” class, which were also difficult to manually categorize, suffered particularly from lack of supporting contextual information. Lack of context at the tweet level is a known challenge for NLP of Twitter text, as users often express complete thoughts over multiple posts. Future research should investigate if incorporating surrounding tweets in the classification model improves overall performance.

For normalization, all systems frequently misclassified closely related concepts (eg, Insomnia and Somnolence) and antonymous concepts (eg, Insomnia and Hypersomnia). For example, in the phrase “sleep for X hours,” only the number of hours spent in sleep can differentiate Hypersomnia (more than 8) from Insomnia (less than 4), and it occurred only twice in the training set, never explicitly mentioning the word night (eg, “waking up in a pool of your own sweat”). Overall, analyses of the errors made by the systems suggest that contextual information is perhaps even more crucial for normalization than classification. The design of the dataset for this subtask does not enable systems to incorporate additional context, and future research should explore the impact of such information.

Table 4. Summary of system extensions and changes in performance compared to the original shared task systems

<table>
<thead>
<tr>
<th>Team</th>
<th>Subtask (evaluation metric)</th>
<th>Extension description</th>
<th>Score</th>
<th>Performance change</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC-Canada</td>
<td>1 (ADR F1-score)</td>
<td>Ensemble of 7 classifiers with random undersampling of the majority class to imbalance ratio of 1:2</td>
<td>0.456</td>
<td>+0.021</td>
</tr>
<tr>
<td>UKNLP</td>
<td>1 (ADR F1-score)</td>
<td>Additional training data, logistic regression and CNN ensembles</td>
<td>0.459</td>
<td>+0.057</td>
</tr>
<tr>
<td>InfyNLP</td>
<td>2 (micro-averaged F1-score for classes 1 and 2)</td>
<td>Additional training data, increased number of random search runs</td>
<td>0.692</td>
<td>-0.001</td>
</tr>
<tr>
<td>NRC-Canada</td>
<td>2 (micro-averaged F1-score for classes 1 and 2)</td>
<td>Additional training data</td>
<td>0.679</td>
<td>+0.0058</td>
</tr>
<tr>
<td>UKNLP</td>
<td>2 (micro-averaged F1-score for classes 1 and 2)</td>
<td>Additional training data (and removed all non-ASCII characters from tweets)</td>
<td>0.694</td>
<td>+0.005</td>
</tr>
<tr>
<td>TurkuNLP</td>
<td>2 (micro-averaged F1-score for classes 1 and 2)</td>
<td>Additional training data</td>
<td>0.665</td>
<td>+0.002</td>
</tr>
<tr>
<td>UKNLP</td>
<td>3 (accuracy)</td>
<td>CNN instead of LSTM at the character level for hierarchical composition</td>
<td>87.7%</td>
<td>+0.5</td>
</tr>
</tbody>
</table>

Figure 4. Distributions of system scores for the 3 subtasks (1, 2, and 3, respectively, from left to right).

Summary of insights gained

The shared task evaluations and post-workshop experiments provided us with insights relevant to future social-media-based text processing tasks beyond the sub-domain of pharmacovigilance. The following list summarizes these insights.

• SVMs, with engineered features and majority-class undersampling, outperformed DNNs for subtask-1. Despite the recent advances in text classification using DNNs, such approaches still underperform for highly imbalanced datasets and may not (yet) be suitable for discovering rare health categories/concepts.

• For tasks with balanced data, DNNs are likely to be more effective than traditional approaches such as SVMs. However, the performances of the different approaches were comparable for subtask-2, and we did not observe any specific set of configurations that performed better.

• Neural-network-based approaches, without requiring any task-specific feature engineering, show low variance in text classification tasks (Figure 5), while SVM performances are very dependent on the feature engineering, weighting, and sampling strategies.
• For text normalization, supervised methods vastly outperform lexicon-based and unsupervised approaches proposed in the past.56
• Large-scale annotation efforts are required to enable systems to accurately identify rare concepts.
• Ensembles of classifiers invariably outperform individual ones, as shown by the post-workshop experiments. However, training and optimizing multiple classifiers, rather than 1, imposes substantial time costs. Therefore, they may be suited only for particularly challenging tasks (eg, subtask-1), where individual classifiers perform significantly worse than human agreement.

Implications for health informatics research
As the volume of health-related data in social media continues to grow, it has become imperative to introduce and evaluate NLP methods that can effectively derive knowledge from it for operational tasks.19,57 Due to the difficulty associated with mining knowledge from social media, earlier approaches primarily attempted to utilize the volume for public health tasks, using keyword-based approaches.58 Text classification is a widely used application of machine learning for extracting information from text, while concept normalization approaches are particularly relevant for social media data due to the necessity of mapping creative expressions to standard forms. While the shared tasks focused on text classification and normalization approaches relevant for the sub-domain of pharmacovigilance, the properties of the texts provided for these tasks are generalizable to many health-related social media tasks. For example, many text classification problems suffer from data imbalance,59 which was a key characteristic of the data for subtask-1. The supervised concept normalization approaches developed by the shared task participants significantly outperformed past efforts, suggesting that our efforts have helped to progress the state of the art in NLP research in this domain. The generalized insights obtained from the large-scale evaluations we reported will serve as guidance, and the public release of the evaluation data with this manuscript will serve as reference standards for future health-related activities.

The progress achieved and the insights obtained through the execution of the shared tasks demonstrate the usefulness of such community-driven developments over publicly released data. We will use the lessons learned to design future shared tasks, such as the inclusion of more contextual information along with the essential texts. Our future efforts will also focus on releasing more health-related annotated datasets from social media.

CONCLUSION
The SMM4H-2017 shared tasks enabled us to advance the current state of NLP methods for mining health-related knowledge from social media texts. We provided training and evaluation data, which exhibited some of the common properties of health-related social media data, for 3 text mining tasks. The public release of the data through the shared tasks enabled the NLP community to participate and evaluate machine learning methods and strategies for optimizing performances on text from this domain. Use of standardized datasets enabled the fast evaluation and ranking of distinct advanced NLP approaches and provided valuable insights regarding the effectiveness of the specific approaches for the given tasks. We have provided a summary of the key findings and lessons learned from the execution of the shared tasks, which will benefit future research attempting to utilize social media big data for health-related activities.

The progress achieved and the insights obtained through the execution of the shared tasks demonstrate the usefulness of such community-driven developments over publicly released data. We will use the lessons learned to design future shared tasks, such as the inclusion of more contextual information along with the essential texts. Our future efforts will also focus on releasing more health-related annotated datasets from social media.

FUNDING
AS and GG were partially supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) grant number NIH NLM R01LM011176. KH, FM, and FG (TurkuNLP) are supported by ATT Tieto käyttöönn grant. SH, TT, AR, and RK (UKNLP) are supported by the NIH National Cancer Institute through grant R21CA218231 and NVIDIA Corporation through the Titan X Pascal GPU donation. MB and GN are supported by the UK EPSRC (grants EP/I028099/1 and EP/N027280/1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Conflict of interest statement. None.

CONTRIBUTORS
AS designed and executed the evaluations for the shared task and drafted the manuscript. GG organized the shared tasks, supervised the evaluations, and contributed to the preparation of the manuscript. SH led participation in tasks-1 and -2 for team UKNLP with assistance in neural modeling from AR, and TT developed the sys-
tem for task-3 with RK guiding the overall participation in terms of both methodology and manuscript writing. KH and FM implemented the CNN for TurkuNLP, including extensions, under the supervision of FG, and all 3 contributed to the preparation of the manuscript. MB implemented the normalization systems for task-3 for gnTeam, and GN provided supervision, and both contributed to the final manuscript. SK led the NRC-Canada efforts, conceived and implemented the system, conducted the experiments, and documented the outcomes. SM and BdB contributed to system conception, and edited the manuscript. JF implemented the CNNs, designed and implemented the system for the InfyNLP team, and performed the experiments. DM supported in performing the experiments and made the primary contribution for the team in drafting the manuscript.

ACKNOWLEDGMENTS
The shared task organizers would like to thank Karen O’Connor and Alexis Upshur for annotating the datasets for the shared task. Computational resources for the TurkuNLP team were provided by CSC - IT Center for Science Ltd, Espoo, Finland. Domain expertise for the University of Manchester team was provided by Professor William G. Dixon, Director of the Arthritis Research U.K. Centre for Epidemiology. The authors would also like to thank the anonymous reviewers for their critiques and suggestions.

REFERENCES


