Journal of the American Medical Informatics Association, 25(10), 2018, 1274–1283
doi: 10.1093/jamia/ocy114
Advance Access Publication Date: 1 October 2018
Research and Applications

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-Hernandez[1]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]School of Computer Science, University of Manchester, Manchester, UK, [3]Infosys Limited, Palo Alto, California, USA, [4]Turku NLP Group, Department of Future Technologies, University of Turku, Turku, Finland, [5]The University of Turku Graduate School, University of Turku, Turku, Finland, [6]Digital Technologies Research Centre, National Research Council Canada, Ottawa, Canada, [7]Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA, [8]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, Kentucky, USA, and [9]Bloomberg, 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 F_1-score) for subtask-1, 0.693 (micro-averaged F_1-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).

Conclusions: Data imbalance and lack of context remain challenges for natural language processing of social media text. Annotated data from the shared task have been made available as reference standards for future studies (http://dx.doi.org/10.17632/rxwfb3tysd.1).

Key words: social media, text mining, natural language processing, pharmacovigilance, machine learning

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,[1] 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.[2] 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,[3] tracking of disease outbreaks,[4,5] charting behavioral factors such as smoking,[6,7] responding to mental health issues,[8,9] and pharmacovigilance.[10] 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.[11] 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 re-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),[12] Text Analysis Conference (TAC),[13] and the International Workshop on Semantic Evaluation (SemEval)[14] have inspired domain-specific counterparts. In the broader medical domain, these include BioASQ,[15] BioCreative,[16] CLEF eHealth,[17] and i2b2,[18] which have significantly advanced health-related NLP.[19]

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.[20] For text classification, data imbalance due to noise and usage of non-standard expressions typically lead to the underperformance of systems[10,21] 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.[22] Within medical NLP, tools utilizing lexicons and knowledge bases such as MetaMap[23] and cTAKES[24] 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.[25]

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[26] 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,[21,27,28] 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 (http://dx.doi.org/10.17632/rxwfb3tysd.1).

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[29]) 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

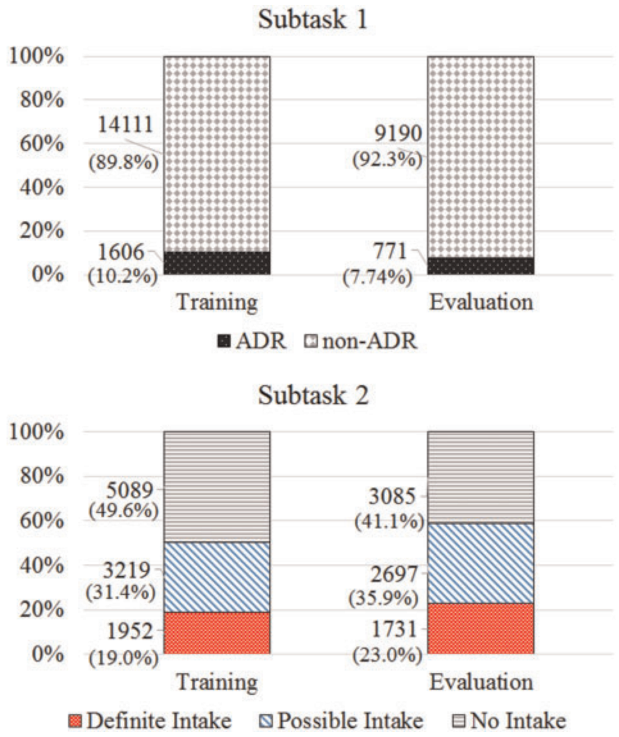


Figure 1. Class distributions for subtasks-1 and -2.

1026 tweets. We double-annotated only a sample of the tweets because of the significant time cost of manual annotation. The annotators followed guidelines that were prepared iteratively until no further improvement in annotation agreement could be achieved.[1] Figure 1 illustrates the distribution of classes over the training and evaluation sets for the 2 subtasks. To ensure that the datasets present the challenges faced by operational systems employed on social media data, we sampled multiple times from a database continuously collecting data. For subtask-1, we used 2 such samples as training data and 1 for evaluation. For subtask-2, to incorporate medication consumption information from a diverse set of users, we drew the training and test sets from distinct users with no overlap.

Training data for subtask-3 consisted of manually curated ADR expressions from tweets mapped to MedDRA[30] (Medical Dictionary for Regulatory Activities) Preferred Terms (PTs). Automatic extraction of ADRs from Twitter has been extensively studied in the recent past, with reported high scores on standard datasets.[31,32] However, the extracted ADRs are often non-standard, creative, or colloquial, and utilizing them for downstream tasks such as signal generation requires normalization, which has been an under-addressed problem. Therefore, we focused on the latter, and we provided pre-extracted ADR expressions with the mappings as input for this subtask. We chose MedDRA as our mapping source because it is specifically designed for documentation and safety monitoring of medicinal products, and is the reference terminology used by regulatory authorities and the pharmaceutical industry for coding ADRs.[33] MedDRA has a hierarchical structure, with Lower Level Terms (LLTs) presenting the most fine-grained level reflecting how an observation might be reported in practice (eg, "*tummy ache*"). Over 70 000 LLTs in this resource are mapped to 22 500 PTs, which represent individual medical concepts such as symptoms (eg, "*abdominal pain*"). The training set consisted of 6650 phrases mapped to 472 PTs (14.09 mentions per concept on average). The evaluation set consisted of 2500 mentions mapped to 254 PTs (9.84 mentions per concept).[2] Figure 2 presents sample instances for the 3 subtasks, along with their manually assigned categories.

Task descriptions and evaluations

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 F_1-score metric, which is based on the true positive (tp), false negative (fn) and false positive (fp) counts:

$$recall = \frac{tp}{tp + fn};precision = \frac{tp}{tp + fp};$$

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

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 F_1-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

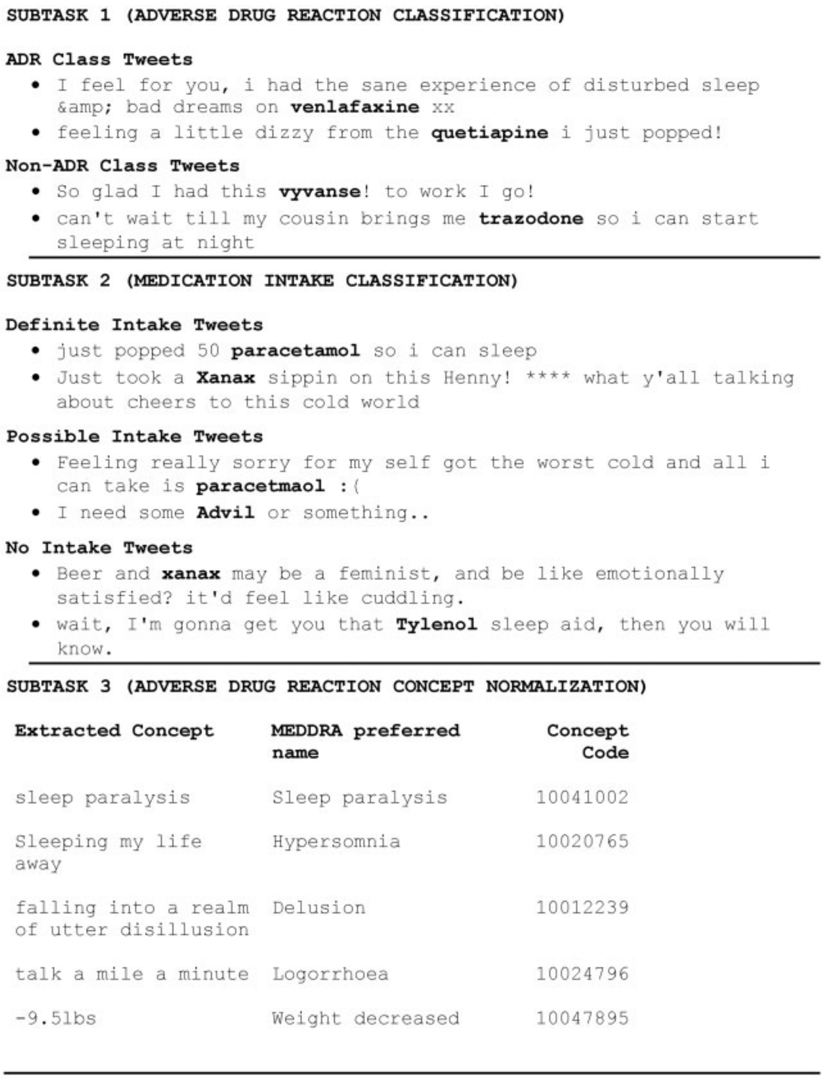
Subtasks-1 and -2: text classification

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[34] along with specialized domain-specific and semantic features. NRC-Canada,[35] the top-performing team for subtask-1, extended its existing state-of-the-art sentiment analysis[36] and stance detection[37] 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[21] are replaced with *<MED>* and *<ADR>*, respectively), pre-trained word embeddings, and word clusters[25] 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,[38] Norms of Valence,

[1] Guidelines are available at http://diego.asu.edu/guidelines/adr_guidelines.pdf (task 1) and https://healthlanguageprocessing.org/twitter-med-intake/ (task 2) (Last accessed: 8/8/2018).

[2] Further details about MedDRA available at: https://www.meddra.org/sites/default/files/guidance/file/intguide_21_0_english.pdf. (Last accessed: 6/7/2018)



SUBTASK 1 (ADVERSE DRUG REACTION CLASSIFICATION)

ADR Class Tweets

- I feel for you, i had the sane experience of disturbed sleep &amp; bad dreams on **venlafaxine** xx
- feeling a little dizzy from the **quetiapine** i just popped!

Non-ADR Class Tweets

- So glad I had this **vyvanse**! to work I go!
- can't wait till my cousin brings me **trazodone** so i can start sleeping at night

SUBTASK 2 (MEDICATION INTAKE CLASSIFICATION)

Definite Intake Tweets

- just popped 50 **paracetamol** so i can sleep
- Just took a **Xanax** sippin on this Henny! **** what y'all talking about cheers to this cold world

Possible Intake Tweets

- Feeling really sorry for my self got the worst cold and all i can take is **paracetmaol** :(
- I need some **Advil** or something..

No Intake Tweets

- Beer and **xanax** may be a feminist, and be like emotionally satisfied? it'd feel like cuddling.
- wait, I'm gonna get you that **Tylenol** sleep aid, then you will know.

SUBTASK 3 (ADVERSE DRUG REACTION CONCEPT NORMALIZATION)

Extracted Concept	**MEDDRA preferred name**	**Concept Code**
sleep paralysis	Sleep paralysis	10041002
Sleeping my life away	Hypersomnia	10020765
falling into a realm of utter disillusion	Delusion	10012239
talk a mile a minute	Logorrhoea	10024796
-9.5lbs	Weight decreased	10047895

Figure 2. Sample instances and their categories for the 3 subtasks. Medication names are shown in bold-face.

Arousal, and Dominance,[39] labMT,[40] and NRC Emoticon Lexicon.[36] The UKNLP (University of Kentucky) systems[41] used similar feature sets (eg, the ADR lexicon) along with 2 additional features: the sum of words' pointwise mutual information (PMI)[42] 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[43] 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.[33,34] The InfyNLP team (Infosys Ltd), top-performers for subtask-2, employed double-stacked ensembles of shallow CNNs.[44] 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[45,46] 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[47] (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 misspellings and convert out-of-vocabulary words to their closest candidates before converting the phrases into dense vector representations using several publicly available sources.[48] 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,[49] 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,[35,41,43,44] 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.

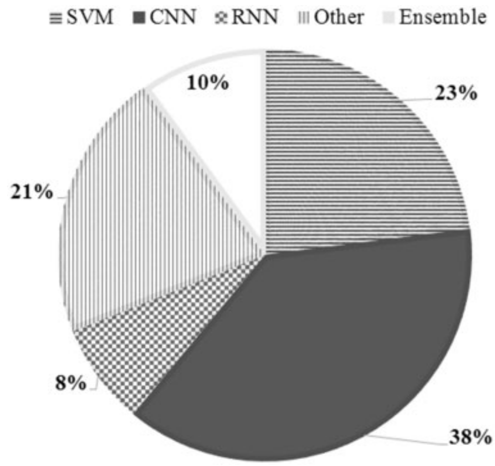


Figure 3. Percentage distributions for 5 categories of approaches attempted by teams for the shared tasks.

Table 1. Performance metrics for selected system submissions for subtask-1, baselines, and system ensembles. Precision, recall, and F_1-score over the ADR class are shown. The top F_1-score among all systems is shown in bold. Detailed discussions about the approaches can be found in the system description papers referenced

System/Team	ADR precision	ADR recall	ADR F_1-score
Baseline 1: Naïve Bayes	0.774	0.098	0.174
Baseline 2: SVMs with RBF kernel	0.501	0.215	0.219
Baseline 3: Random Forest	0.429	0.066	0.115
NRC-Canada[35]	0.392	0.488	0.435
CSaRUS-CNN[50] (Arizona State University)	0.437	0.393	0.414
NorthEasternNLP[51] (NorthEastern University)	0.395	0.431	0.412
UKNLP[41] (University of Kentucky)	0.498	0.337	0.402
TsuiLab[52] (University of Pittsburgh)	0.336	0.348	0.342
Ensemble all: best configuration (>6 ADR votes)	0.435	0.492	0.461
Ensemble top 7: majority vote (>3)	0.529	0.398	0.454
Ensemble top 7: >2 ADR votes	0.462	0.492	**0.476**
Ensemble top 5: majority vote (>2)	0.521	0.415	0.462
Ensemble top 5: at least 1 ADR vote	0.304	0.641	0.413
Ensemble top 3: >1 ADR vote	0.464	0.441	0.452

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 F_1-score of 0.435. InfyNLP's ensemble of shallow CNNs topped subtask-2 with a micro-averaged F_1-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 2. Performance metrics for selected system submissions for subtask-2, baselines, and system ensembles. Micro-averaged precision, recall, and F_1-scores are shown for the *definite intake* (class 1) and *possible intake* (class 2) classes. The highest F_1-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)

System/Team	Micro-averaged precision for classes 1 and 2	Micro-averaged recall for classes 1 and 2	Micro-averaged F_1-score for classes 1 and 2
Baseline 1: Naïve Bayes	0.359	0.503	0.419
Baseline 2: SVMs	0.652	0.436	0.523
Baseline 3: Random Forest	0.628	0.487	0.549
InfyNLP[44] (Infosys Ltd)	0.725	0.664	0.693
UKNLP[41] (University of Kentucky)	0.701	0.677	0.689
NRC-Canada[35]	0.708	0.642	0.673
TJIIP (Tongji University, China)	0.691	0.641	0.665
TurkuNLP[43] (University of Turku)	0.701	0.630	0.663
CSaRUS-CNN[50] (Arizona State University)	0.709	0.604	0.652
NTTMU[53] (Multiple Universities, Taiwan)	0.690	0.554	0.614
Ensemble all: majority vote	0.736	0.657	0.694
Ensemble top 10: majority vote	0.726	0.679	**0.702**
Ensemble top 7: majority vote	0.724	0.673	0.697
Ensemble top 5: majority vote	0.723	0.667	0.694
Ensemble top submissions from top 5 teams: majority vote	0.727	0.673	0.699

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

Team	Approach summary	Accuracy (%)
Baseline 1	Exact lexical match with MedDRA PT	11.6
Baseline 2	Exact lexical match with MedDRA LLT or PT	25.1
Baseline 3	Match with training set annotation	63.5
gnTeam[54] (University of Manchester)	Multinomial Logistic Regression	87.7
	RNN with GRU	85.5
	Ensemble	88.5
UKNLP[41] (University of Kentucky)	Hierarchical RNN with LSTM	87.2
	Hierarchical RNN with LSTM and external data	86.7
Ensemble	All systems	**88.7**
	Top 3	**88.7**

paper and associated system descriptions.[26,35,41,43,44,54] Figure 4 illustrates the distributions of all the individual system scores.

Tables 1–3 illustrate that for all 3 subtasks, some combination of system ensembles outperform the top system. For subtask-1, the best ADR F_1-score (0.476) on the test dataset was obtained by taking the top 7 systems and using a voting threshold of 2 (Table 1). For subtask-2, majority voting from the top 10 systems obtained the highest F_1-score (0.702). For subtask-3, both ensembles outperformed the individual submissions (accuracy = 88.7%), albeit marginally.

Post-workshop follow-up modifications

Both the UKNLP and NRC-Canada teams were able to marginally improve the performances of their systems for subtask-1 by using additional data or by modifying their systems. The NRC-Canada team reported that ensembles of 7 to 9 classifiers, each trained on a random sub-sample of the majority class to reduce class imbalance to 1: 2, outperformed their top-performing system. The UKNLP team reported that the additional training data improved the performance of their logistic regression classifier for the task, which consequently improved the performance of the logistic regression and CNN ensembles, increasing the best ADR F_1-score to 0.459 (+0.057).

For subtask-2, NRC-Canada reported that domain-generalized n-grams showed significant increases in performance, while sentiment lexicons were not useful. For CNN-based systems (eg, UKNLP, TurkuNLP, and InfyNLP), incorporation of additional training data showed slight improvements in performances. Only UKNLP attempted a system extension for subtask-3, and they slightly improved accuracy by employing a CNN instead of an LSTM at the character level for the hierarchical composition. None of these system extensions performed better than the multi-system ensembles presented in Tables 1–3. Table 4 summarizes the system extensions and their performances.

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-

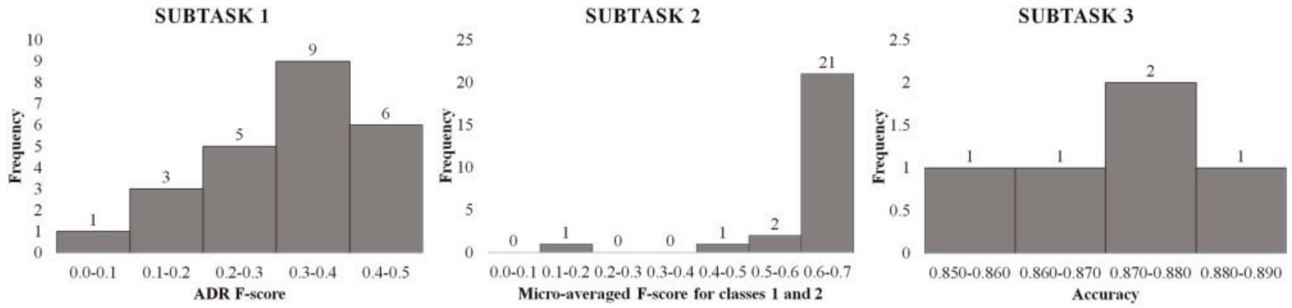


Figure 4. Distributions of system scores for the 3 subtasks (1, 2, and 3, respectively, from left to right).

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

Team	Subtask (evaluation metric)	Extension description	Score	Performance change
NRC-Canada	1 (ADR F_1-score)	Ensemble of 7 classifiers with random undersampling of the majority class to imbalance ratio of 1: 2	0.456	+0.021
UKNLP	1 (ADR F_1-score)	Additional training data, logistic regression and CNN ensembles	0.459	+0.057
InfyNLP	2 (micro-averaged F_1-score for classes 1 and 2)	Additional training data, increased number of random search runs	0.692	−0.001
NRC-Canada	2 (micro-averaged F_1-score for classes 1 and 2)	Additional training data	0.679	+0.0058
UKNLP	2 (micro-averaged F_1-score for classes 1 and 2)	Additional training data (and removed all non-ASCII characters from tweets)	0.694	+0.005
TurkuNLP	2 (micro-averaged F_1-score for classes 1 and 2)	Additional training data	0.665	+0.002
UKNLP	3 (accuracy)	CNN instead of LSTM at the character level for hierarchical composition	87.7%	+0.5

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 is challenging to incorporate this knowledge into the machine learning models. Lack of training data for rarely occurring concepts was another cause of errors. For example, the concept "*Night sweats*" was frequently misclassified (usually as "*Hyperhidrosis*"), 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.

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,[55] 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.

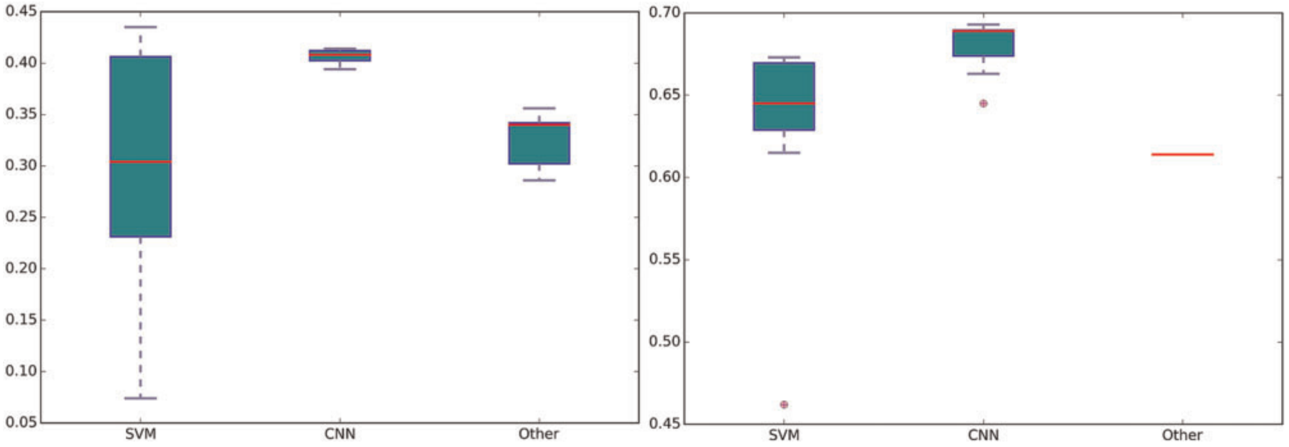


Figure 5. Boxplots illustrating the performances of *SVMs*, *CNNs*, and *Other* classification strategies for subtasks-1 and -2.

- 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 studies 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 käyttöön 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 system 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 expertize 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

1. PEW Research Center. Demographics of Social Media Users and Adoption in the United States. 2017. http://www.pewinternet.org/fact-sheet/social-media/. Accessed March 3, 2018.
2. Kennedy B, Funk C. *Public Interest in Science and Health Linked to Gender, Age and Personality.* PEW Research Center; 2015. http://www.pewinternet.org/2015/12/11/public-interest-in-science-and-health-linked-to-gender-age-and-personality/. Accessed July 1, 2018.
3. Paul MJ, Dredze M. You are what you Tweet: analyzing Twitter for public health. *Proc Fifth Int AAAI Conf Weblogs Soc Media* 2011; 265–72. doi: 10.1.1.224.9974.
4. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Sroudsburg, PA: Association for Computational Linguistics; 2011: 1568–1576.
5. Mollema L, Harmsen IA, Broekhuizen E, *et al*. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in the Netherlands in 2013. *J Med Internet Res* 2015; 17 (5): e128.
6. Aphinyanaphongs Y, Lulejian A, Brown DP, Bonneau R, Krebs P. Text Classification for Automatic Detection of E-cigarette Use and Use for Smoking Cessation from Twitter: A Feasibility Pilot. *Pac Symp Biocomput*. 2016;21:480–91. Singapore: World Scientific Publishing Company. doi: 10.1142/9789814749411_0044
7. Struik LL, Baskerville NB. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis of posts. *J Med Internet Res* 2014; 16 (7): e170.
8. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting changes in suicide content manifested in social media following celebrity suicides. In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media—HT '15*. New York, NY: ACM Press; 2015: 85–94.
9. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics*. Denver, CO. 2015: 31–39.
10. Sarker A, Ginn R, Nikfarjam A, *et al*. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform* 2015; 54: 202–12.
11. Brennan PF, Chiang MF, Ohno-Machado L. Biomedical informatics and data science: evolving fields with significant overlap. *J Am Med Inform Assoc*. 2018; 25 (1): 2–3.
12. The SIGNLL Conference on Computational Natural Language Learning. 2017. http://www.conll.org/. Accessed July 12, 2017.
13. National Institute of Standards and Technology. Text Analysis Conference. 2017. https://tac.nist.gov/. Accessed July 12, 2017.
14. International Workshop on Semantic Evaluation. 2017. http://alt.qcri.org/semeval2018/. Accessed July 12, 2017.
15. BioASQ. 2017. http://www.bioasq.org/. Accessed July 12, 2017.
16. BioCreative. 2017. http://www.biocreative.org/. Accessed July 12, 2017.
17. CLEF eHealth 2018. Lab Overview CLEF eHealth. 2018. https://sites.google.com/view/clef-ehealth-2018/. Accessed July 12, 2017.
18. i2b2 - Informatics for Integrating Biology and the Bedside. https://www.i2b2.org/. Accessed August 2, 2018.
19. Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearb Med Inform* 2016; 25 (1): 224–33.
20. Han B, Cook P, Baldwin T. Lexical normalization for social media text. *ACM Trans Intell Syst Technol* 2013; 4(1): Article No. 5, 1–27.
21. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform* 2015; 53: 196–207.
22. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearb Med Inform* 2017; 26 (01): 214–27.
23. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17 (3): 229–36.
24. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
25. 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 Informatics Assoc* 2015; 22 (3): 671–81.
26. Sarker A, Gonzalez-Hernandez G. Overview of the Second Social Media Mining for Health (SMM4H) shared tasks at AMIA 2017. In: *Proceedings of the 2nd Social Media Mining for Health Research and Applications Workshop Co-Located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 43–48. http://ceur-ws.org/Vol-1996/paper8.pdf. Accessed December 6, 2017.
27. Klein A, Sarker A, Rouhizadeh M, O'Connor K, Gonzalez G. Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system. In: *Proceedings of the BioNLP 2017 Workshop*. Vancouver, BC, Canada: Association for Computational Linguistics; 136–142.
28. Sarker A, Nikfarjam A, Gonzalez G. Social Media Mining Shared Task Workshop. *Pac Symp Biocomput. World Scientific Publishing Company, Singapore*. 2016; 21: 581–592.
29. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20 (1): 37–46.
30. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf*. 1999 Feb; 20 (2): 109–17.
31. O'Connor K, Nikfarjam A, Ginn R, *et al*. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. *AMIA Annu Symp Proc*. 2014; 2014: 924–933.
32. 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 Inform Assoc*. 2017; 24 (4): 813–821.
33. Souvignet J, Declerck G, Asfari H, Jaulent MC, Bousquet C. OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. *J Biomed Inform*. 2016; 63: 100–107.
34. Owoputi O, O'Connor B, Dyer C, Gimpel K, Schneider N, Smith NA. Improved part-of-speech tagging for online conversational text with word clusters. In: *Proceedings of the 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA: Association for Computational Linguistics; 2013: 380–390.

35. Kiritchenko S, Mohammad SM, Morin J, de Bruijn B. NRC-Canada at SMM4H shared task: classifying Tweets mentioning adverse drug reactions and medication intake. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 1–11. http://ceur-ws.org/Vol-1996/paper1.pdf. Accessed December 6, 2017.
36. Kiritchenko S, Zhu X, Mohammad SM. Sentiment of short informal texts. *J Artif Intell Res* 2014; 50 (1): 723–62. https://dl.acm.org/citation.cfm?id=2693087.
37. Mohammad SM, Sobhani P, Kiritchenko S. Stance and sentiment in tweets. *ACM Trans Internet Technol* 2017; 17 (3): 1–23.
38. Hu M, Liu B. Mining and summarizing customer reviews. *Proc 2004 ACM SIGKDD Int Conf Knowl Discov Data Min KDD 04* 2004; 4: 168.
39. Warriner AB, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13, 915 English lemmas. *Behav Res Methods* 2013; 45 (4): 1191–207.
40. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One* 2011; 6 (12): e26752.
41. Han S, Tran T, Rios A, Kavuluru R. Team UKNLP: detecting ADR Mentions on Twitter. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 49–53. http://ceur-ws.org/Vol-1996/paper9.pdf. Accessed April 30, 2018.
42. Bouma G. Normalized (pointwise) mutual information in collocation extraction. *Proc Ger Soc Comput Linguist (GSCL 2009)* 2009; 31–40.
43. Hakala K, Mehryary F, Moen H, Kaewphan S, Salakoski T, Ginter F. Ensemble of convolutional neural networks for medicine intake recognition in Twitter. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 59–63. http://ceur-ws.org/Vol-1996/paper11.pdf. Accessed April 30, 2018.
44. Friedrichs J, Mahata D, Gupta S. InfyNLP at SMM4H task 2: stacked ensemble of shallow convolutional neural networks for identifying personal medication intake from Twitter. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 68–71. http://ceur-ws.org/Vol-1996/paper13.pdf. Accessed April 30, 2018.
45. Godin F, Vandersmissen B, De Neve W, Van de Walle R. Multimedia Lab @ ACL W-NUT NER shared task: named entity recognition for Twitter microposts using distributed word representations. In: *Workshop on Noisy User-Generated Text, ACL 2015*. Beijing, China: Association for Computational Linguistics; 2015: 146–153. doi: 10.1126/science.1247727.
46. Shin B, Lee T, Choi JD. Lexicon integrated CNN models with attention for sentiment analysis In: *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen; 2017: 149–158. http://www.aclweb.org/anthology/W17-5220. Accessed March 5, 2018.
47. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–57. https://www.jair.org/media/953/live-953-2037-jair.pdf. Accessed March 5, 2018.
48. Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: language models and their utilities. *Data Br* 2017; 10: 122–31.
49. Jozefowicz R, Zaremba W, Sutskever I. An Empirical Exploration of Recurrent Network Architectures. http://proceedings.mlr.press/v37/jozefowicz15.pdf?utm_campaign=Revue newsletter&utm_medium=Newsletter&utm_source=revue. Accessed February 20, 2018.
50. Magge A, Scotch M, Gonzalez G. CSaRUS-CNN at AMIA-2017 tasks 1, 2: under sampled CNN for text classification. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 76–78. http://ceur-ws.org/Vol-1996/paper15.pdf. Accessed May 8, 2018.
51. Jain S, Peng X, Wallace BC. Detecting Twitter posts with adverse drug reactions using convolutional neural networks. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 72–75. http://ceur-ws.org/Vol-1996/paper14.pdf. Accessed May 8, 2018.
52. Tsui F, Shi L, Ruiz V, *et al*. Detection of adverse drug reaction from Twitter data. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 64–67. http://ceur-ws.org/Vol-1996/paper12.pdf. Accessed May 8, 2018.
53. Wang C-K, Chang N-W, Su EC, Dai H-J. NTTMU system in the 2nd social media mining for health applications shared task. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 83–86. http://ceur-ws.org/Vol-1996/paper17.pdf. Accessed May 8, 2018.
54. Belousov M, Dixon W, Nenadic G. Using an ensemble of linear and deep learning models in the SMM4H 2017 medical concept normalization task. In: *Proceedings of the Second Workshop on Social Media Mining for Health Research and Applications Workshop Co-located with the American Medical Informatics Association Annual Symposium (AMIA 2017)*; 2017: 54–58. http://ceur-ws.org/Vol-1996/paper10.pdf. Accessed May 8, 2018.
55. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha; 2014: 1746–1751. http://www.aclweb.org/anthology/D14-1181. Accessed March 8, 2018.
56. Emadzadeh E, Sarker A, Nikfarjam A, Gonzalez G. Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. *AMIA Annu Symp Proc*. Washington, DC; 2017: 679–688.
57. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform* 2015; 10 (1): 183–93. Stuttgart, Germany: Georg Thieme Verlag KG.
58. Charles-Smith LE, Reynolds TL, Cameron MA, *et al*. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. Braunstein LA, ed. *PLoS One* 2015; 10 (10): e0139701.
59. Li Y, Sun G, Zhu Y. Data imbalance problem in text classification. In: *2010 Third International Symposium on Information Processing*. IEEE; 2010: 301–305. doi: 10.1109/ISIP.2010.47.