Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017

Abeed Sarker, Ph.D.[1], Graciela Gonzalez-Hernandez, Ph.D.[2]
[1]Health Language Processing Laboratory, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Abstract

The volume of data encapsulated within social media continues to grow, and, consequently, there is a growing interest in developing effective systems that can convert this data into usable knowledge. Over recent years, initiatives have been taken to enable and promote the utilization of knowledge derived from social media to perform health related tasks. These initiatives include the development of data mining systems and the preparation of datasets that can be used to train such systems. The overarching focus of the SMM4H shared tasks is to release annotated social media based health related datasets to the research community, and to compare the performances of distinct natural language processing and machine learning systems on tasks involving these datasets. The second execution of the SMM4H shared tasks comprised of three subtasks involving annotated user posts from Twitter (tweets): (i) automatic classification of tweets mentioning an adverse drug reaction (ADR) (ii) automatic classification of tweets containing reports of first-person medication intake, and (iii) automatic normalization of ADR mentions to MedDRA concepts. A total of 15 teams participated and 55 system runs were submitted. The best performing systems for tasks 2 and 3 outperformed the current state of the art systems.

Introduction

The second execution of the SMM4H shared tasks built on the success of the first execution of the shared task workshop[1], which was held at the Pacific Symposium on Biocomputing (PSB), 2016. In line with the previous shared task, the data comprised of medication mentioning posts from Twitter, which were retrieved using the Twitter public streaming API[2]. We designed and provided annotated data for three tasks. The annotated data were made publicly available for download. The performances of participating systems were compared on blind evaluation sets for each task.

Shared Task Design

The overall shared task consisted of three independent tasks/subtasks. Teams could participate in one or multiple tasks. From the perspective of text mining, the first two tasks focused on text classification and the third task focused on concept normalization. Manually annotated training data for the three tasks were made available to the participants in May, 2016. Unlabeled evaluation data was released in September, 2016. Evaluations of participant submissions were conducted from 5th to 12th September. In total, 15 teams participated in the shared tasks and 55 system runs were accepted from them (maximum of three submissions per team per task). We received 24 submissions for task 1, 26 for task 2 and 5 for task 3. Participating teams were invited to submit system descriptions to describe their approaches to the tasks. Teams participating in multiple tasks submitted a single system description. Each system description was peer reviewed by at least one reviewer. Nine system descriptions were accepted for inclusion in the SMM4H workshop proceedings, including one system description that was accepted as a full paper at the workshop after undergoing peer review by two reviewers. We provide descriptions of the three tasks and the associated data in the following sections/subsections.

Task Descriptions

Tasks

The primary goal of the SMM4H shared tasks is to promote community driven development and evaluations of systems focusing on social media based health data. This year's tasks involved medication-mentioning user posts from Twitter. We included two tasks from the last execution at PSB and a new task. Outlines of the tasks are as follows:

(i) Automatic classification of ADR mentioning tweets. This is a binary text classification task for which systems were required to predict if a tweet mentions an ADR or not. Such a system is crucial for active surveillance of ADRs from social media data as most of the medication-related chatter in the domain, including those on Twitter, are noise. This task was also part of the first execution of the SMM4H shared tasks. Further details about this task can be found in our past publication[3].

(ii) Automatic classification of medication intake mentioning posts. This is a three-class text classification task. Each medication-mentioning tweet is categorized into three classes—*definite intake* (where the user presents clear evidence of personal consumption), *possible intake* (where it is likely that the user consumed the medication, but the evidence is unclear), and *no intake* (where there is no evidence that the user consumed the medication). This proposed task was new in the 2017 SMM4H shared tasks. Further details about this task can be found in our recent publication[4].

(iii) Normalization of ADR mentions. The goal of this task is to normalize different natural language expressions of the same ADR concept into standard IDs. This is a particularly challenging task and although it was proposed in the first execution of the shared tasks, there were no participants.

To facilitate the shared task, we made available large annotated Twitter data sets. The overall shared task was designed to capitalize on the interest in social media mining and appeal to a diverse set of researchers working on distinct topics such as natural language processing, biomedical informatics, and machine learning. The different subtasks presented a number of interesting challenges including the noisy nature of the data, the informal language of the user posts, misspellings, and data imbalance. We provide details of the data used for each of the three abovementioned tasks, and the tasks themselves, in the following subsection.

Data

The dataset made available for the shared tasks were collected from Twitter using the public streaming API. The annotated datasets provided as training sets were made available to the public with our prior publications[3,4]. Only task 3 included new, previously unpublished data for training.

Task 1: ADR Classification. Participants were provided with the training/development set containing tweets which were annotated in a binary fashion to indicate the presence or absence of ADRs. Initially, a total of 10,822 annotated tweets were made available[1]. Later on, an additional 4895 tweets were released in the same fashion to active participants (previous shared task's evaluation set). The evaluation set consisted of 9961 tweets. The per-class distributions of the tweets in the three sets are shown in Table 1. The evaluation metric for this task was the F-score for the ADR class, since the primary intent of this task is to be able to filter out ADR indicating tweets from large amounts of noise.

Table 1. Training and evaluation datasets for task 1 of the SMM4H shared tasks.

Set	Total Number of Tweets	Number of Tweets in ADR Class	Number of Tweets in non-ADR Class
Training 1	10,822	1239	9583
Training 2	4895	367	4528
Evaluation	9961	771	9190

Task 2: Medication Intake Classification. Participants were provided with tweets that have been manually categorized into three classes—*definite intake*, *possible intake* and *no intake.* Like task 1, data was released in three phases. Initially, 8000 annotated tweets were released, followed by an additional 2260 tweets for active participants. The evaluation set consisted of 7513 tweets. The per-class distributions of the tweets are shown in Table 2. For this task, the evaluation metric was micro-averaged F-score for the definite intake and possible intake classes. This metric was chosen for evaluation because the tweets belonging to these two classes are of interest in social media based drug safety surveillance systems, while the *no intake* class primarily represents noise.

[1] Due to Twitter's privacy policy, the actual tweets were not shared publicly. We made available a download script and the TweetIDs and UserIDs for the tweets. The publicly available tweets can be downloaded using the download scripts.

Table 2. Training and evaluation datasets for task 2 of the SMM4H shared tasks.

Set	Total Number of Tweets	Number of Tweets in the *Definite Intake* Class	Number of Tweets in the *Possible Intake* Class	Number of Tweets in the *No Intake* Class
Training 1	8000	1528	2502	3970
Training 2	2260	424	717	1119
Evaluation	7513	1731	2697	3085

Task 3: Adverse Drug Reaction Mention Normalization. The training data consisted of ADR mentions mapped to MedDRA (Medical Dictionary for Regulatory Activities)[3] Preferred Terms (PTs). The training set consisted of 6,650 phrases mapped to 472 PTs (14.09 mentions per concept on average). The test set consisted of 2500 mentions mapped to 254 classes. The evaluation metric for this task was accuracy (*i.e.*, number of correctly identified MedDRA PTs divided by the total number of instances in the evaluation set).

Results

Task 1

Eleven teams registered to participate in the task and 24 submissions from nine teams were included in the final evaluations. System submissions were excluded if they did not meet the deadline, were incompatible, did not follow the shared task guidelines or were incomplete. Table 3 presents the performances of the 24 included systems grouped by the team names. Team NRC_Canada had the best performing system at for this task, obtaining an ADR class F-score of 0.435[5].

Table 3. System performances for each team for task 1 of the shared task. Precision, recall and F-score over the ADR class is shown. Top score in each column is shown in bold.

Team	Institution(s) - Country	ADR Precision	ADR Recall	ADR F-score
TsuiLab	University of Pittsburgh – United States	0.333	0.350	0.341
		0.298	0.394	0.339
		0.336	0.348	0.342
NRC_Canada	National Research Council – Canada	0.392	**0.488**	**0.435**
		0.386	0.413	0.399
		0.464	0.396	0.427
NorthEasternNLP	Northeastern University – United States	0.551	0.306	0.394
		0.395	0.431	0.412
NTTMU	Taipei Medical University, Academia Sinica, National Taitung University – Taiwan	0.213	0.433	0.286
		0.362	0.249	0.295
		0.226	0.403	0.290
CSaRUS-CNN	Arizona State University – United States	0.437	0.393	0.414
		0.467	0.357	0.404
		0.396	0.431	0.412
TJIIP	University of Montreal – Canada	0.359	0.398	0.378
		0.422	0.154	0.226

[3] Available at: https://www.meddra.org/.

		0.325	0.400	0.359
UKNLP	University of Kentucky – United States	0.459	0.237	0.313
		0.567	0.259	0.356
		0.498	0.337	0.402
deepCyberNet	Amrita School of Engineering Coimbatore – India	0.078	0.170	0.107
AMRITA_CEN_ NLP_RBG	Amrita School of Engineering Coimbatore – India	0.056	0.109	0.074
		0.087	0.204	0.121
		0.186	0.481	0.268

Task 2

Eleven teams registered to participate in this task including eight teams that also registered for Task 1. 26 submissions from ten teams were included in the final evaluations. Exclusion criteria were identical to those of task 1. Table 2 presents the performances of these 26 systems grouped by team names. Team InfyNLP had the best performing system for this task, obtaining micro-averaged F-score of 0.693 for the two relevant classes[6].

Table 4. System performances for each team for task 2 of the shared task. Micro-averaged precision, recall and F-scores are shown for the *definite intake* (class 1) and *possible intake* (class 2) classes. Top score in each column is shown in bold.

Team	**Institution(s) – Country**	**Micro-averaged precision for classes 1 and 2**	**Micro-averaged recall for classes 1 and 2**	**Micro-averaged F-score for classes 1 and 2**
CSaRUS-CNN	Arizona State University – United States	0.696	0.601	0.645
		0.708	0.599	0.649
		0.709	0.604	0.652
AMRITA_CEN_ NLP_RBG	Amrita School of Engineering Coimbatore – India	0.569	0.390	0.462
NRC_Canada	National Research Council – Canada	0.708	0.642	0.673
		0.705	0.639	0.671
		0.704	0.635	0.668
NTTMU	Taipei Medical University, Academia Sinica, National Taitung University – Taiwan	0.690	0.554	0.614
		0.644	0.588	0.615
		0.662	0.572	0.614
RITUAL	University of Houston – United States	0.630	0.571	0.599
		0.643	0.578	0.609
		0.650	0.575	0.610
TJIIP	University of Montreal - Canada	0.691	0.641	0.665
		0.628	0.557	0.590
		0.654	0.664	0.659
TurkuNLP		0.692	0.601	0.643

	University of Turku, Turku Centre for Computer Science – Finland	0.701	0.630	0.663
UKNLP	University of Kentucky – United States	0.688	0.607	0.645
		0.705	0.666	0.685
		0.701	**0.677**	0.689
InfyNLP	Infosys Ltd. – United States Indian Institute of Technology – India	0.716	0.664	0.689
		0.721	0.661	0.690
		0.725	0.664	**0.693**
deepCyberNet	Amrita School of Engineering Coimbatore – India	0.414	0.107	0.171
		0.843	0.487	0.617

Task 3

Two teams registered to participate in this task and five system submissions were submitted. Table 5 summarizes the performances of the five systems. It can be seen from the table that the different systems showed similar performances, with one system from team gnTeam obtaining the best accuracy of 88.5%[7].

Table 5. System performances for task 3. Accuracies over the evaluation set are shown. Best performance is shown in bold.

Team	Institution – Country	Accuracy (%)
gnTeam	University of Manchester – United Kingdom	87.7
		85.5
		88.5
UKNLP	University of Kentucky – United States	87.2
		86.7

Conclusion

The number of submissions received for the second execution of the SMM4H shared tasks was more than double of that received for the first execution. The submitted systems employed a wide range of machine learning methods. The system descriptions that have been published with the shared task proceedings provide further details about these methods and the relative performances of each. The successful execution of the shared tasks suggests that this is an effective model for encouraging community-driven development of systems for social media based heath related text mining, and warrants further future efforts.

Acknowledgments

This work was supported by National Institutes of Health (NIH) National Library of Medicine (NLM) grant number NIH NLM 5R01LM011176. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM or NIH.

The authors would like to thank the members of the Health Language Processing Laboratory for their support. In particular, the authors would like to acknowledge the efforts by Karen O'Connor and Alexis Upshur for preparing the annotated data sets. The authors would also like to thank Dr. Davy Weissenbacher, Dr. Masoud Rouhizadeh and Dr. Ari Z. Klein for reviewing system descriptions and contributing to the selection process.

References

1. Sarker A, Nikfarjam A, Gonzalez G. SOCIAL MEDIA MINING SHARED TASK WORKSHOP. *Pac Symp Biocomput*. 2016;21:581-592. http://www.ncbi.nlm.nih.gov/pubmed/26776221. Accessed January 5, 2017.

2. Twitter. Twitter Public Streaming API. https://developer.twitter.com/en/docs.
3. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform*. 2014;53:196-207. doi:10.1016/j.jbi.2014.11.002.
4. 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; :136-142.
5. 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 Applications (SMM4H)*. Health Language Processing Laboratory; 2017.
6. 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 Applications (SMM4H)*. Health Language Processing Laboratory; 2017.
7. Belousov M, Dixon W, Nenadic G. Using an ensemble of linear and deep learning models in the SMM4H 2017 medical concept normalisation task. In: *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H)*. Health Language Processing Laboratory; 2017.