Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018

Davy Weissenbacher[†], **Abeed Sarker**[†], **Michael Paul**[‡], **Graciela Gonzalez-Hernandez**[†]
[†]DBEI, Perelman School of Medicine, University of Pennsylvania, PA, USA
[‡]Information Science University of Colorado Boulder, CO, USA
{dweissen,abeed,gragon}@pennmedicine.upenn.edu
mpaul@colorado.edu

Abstract

The goals of the SMM4H shared tasks are to release annotated social media based health related datasets to the research community, and to compare the performances of natural language processing and machine learning systems on tasks involving these datasets. The third execution of the SMM4H shared tasks, co-hosted with EMNLP-2018, comprised of four subtasks. These subtasks involve annotated user posts from Twitter (tweets) and focus on the (i) automatic classification of tweets mentioning a drug name, (ii) automatic classification of tweets containing reports of first-person medication intake, (iii) automatic classification of tweets presenting self-reports of adverse drug reaction (ADR) detection, and (iv) automatic classification of vaccine behavior mentions in tweets. A total of 14 teams participated and 78 system runs were submitted (23 for task 1, 20 for task 2, 18 for task 3, 17 for task 4).

1 Introduction

The third execution of the SMM4H shared tasks built on the success of the two previous shared task workshops, which were held at the Pacific Symposium on Biocomputing (PSB) in 2016 and at the AMIA Annual Symposium in 2017. In line with the previous shared tasks, the data comprised of medication mentioning posts from Twitter, which were retrieved using the Twitter public streaming API. For this iteration, We designed and provided annotated data for four tasks. The annotated data were made publicly available for download. The performances of participating systems were compared on blind evaluation sets for each task.

1.1 Shared Task Design

Teams were allowed to participate in one or multiple tasks. In order to analyze cross-task application of classification techniques, all the tasks for this year's execution focused on text classification. Manually annotated training data for the four tasks were made available to the participants in May, 2018. Unlabeled evaluation data was released in July, 2018. Evaluations of participant submissions were conducted from 29th July to 2nd of August. In total, 14 teams participated in the shared tasks and 78 system runs were accepted from them (maximum of three submissions per team per task). We received 23 submissions for task 1, 20 for task 2, 18 for task 3, 17 for subtask 4. Participating teams were invited to submit system descriptions to describe their approaches to the tasks. We provide descriptions of the four tasks and the associated data in the following sections/subsections.

2 Task Descriptions

2.1 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 AMIA and two new task. Outlines of the tasks are as follows:

1. Automatic classification of posts mentioning a drug name. In this binary classification task, the systems were required to distinguish tweets mentioning any drug names or dietary supplement. Often run first in automatic pipelines mining health related information in social media, the performances obtained on this task conditions the overall performances of the entire pipelines and their usefulness. This proposed task was new and intended to establish common baselines for future research.

2. Automatic classification of medication intake mentioning posts. This is a three-class

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 13–16
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

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).

3. 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 does not represent ADRs.

4. Automatic classification of vaccine behavior mentions in tweets. Specifically, English-language tweets are classified to indicate whether the user intends to receive a seasonal influenza (flu) vaccine (Huang et al., 2017). It is a binary classification task where the positive class indicates that the user has received or intends to receive the current flu vaccine, and all other tweets (which are filtered with vaccine-related keywords) are labeled negative. Such a classifier can be used to measure patterns in vaccination behaviors across populations.

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 four above-mentioned tasks, and the tasks themselves, in the following subsections.

2.2 Data

The dataset made available for the shared tasks were collected from Twitter using the public streaming API. Task 1 and task 4 included new and unpublished annotated datasets provided as training and testing sets. Tasks 2 and 3 re-used existing training datasets from the SMM4H-2017 shared tasks where the SMM4H-2017 shared tasks' evaluation sets were included in the training datasets used this year. These datasets had been made available with our prior publication following the execution of the past workshop (Sarker et al., 2018).

Task 1: Drug names detection. Participants were given tweets with binary annotation, indicating the presence or absence in the tweet of one or more drug names/dietary supplement, manually created. The data were released in two phases. An initial set of 9,622 tweets were made available[1] for training to any participants. The test set composed of 5,382 tweets was distributed only to registered participants. Both training and test sets were balanced with 4647/2530 tweets mentioning no drug and 4975/2852 tweets mentioning at least one drug, respectively. All participants were evaluated using common metrics for binary classification: Precision, Recall and F-score for tweets mentioning a drug.

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. Data was released in the same manner as task 1. 17,773 annotated tweets were made available for training. The evaluation set consisted of 5000 tweets. 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.

Task 3: 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. A total of 25,633 annotated tweets were made available for training. The evaluation set consisted of 5000 tweets. 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.

Task 4: Vaccine Behavior Classification. Par-

[1] Due to Twitter's privacy policy, the actual tweets were not shared publicly. We made available the TweetIDs and UserIDs for the tweets as well as a download script to download all publicly available tweets.

Team	Institution(s)-Country	P	R	F
ART	Tata Consultancy Services Limited, India	0.785	0.880	0.830
CIC-NLP	Instituto Politecnico Nacional, Mexico	0.920	0.899	0.910
ClaC	Concordia University, Canada	0.788	0.769	0.778
IIT_KGP	Indian Institute of Technology, India	0.918	0.840	0.877
IRISA	INRIA-IRISA, France	0.922	0.906	0.914
LILU	Technical University of Moldova, Moldova	0.841	0.860	0.850
Techno	University Abou Bekr Belkaid, Algeria	0.905	0.855	0.879
THU_NGN	Tsinghua University, China	0.933	0.904	**0.918**
Tub-Oslo	University of Tubingen, Germany University of Oslo, Norway	0.917	**0.907**	0.912
UChicagoCompLx	University of Chicago, USA	**0.937**	0.891	0.914
UZH	University of Zurich, Switzerland	0.927	0.878	0.902

Table 1: System performances for each team for task 1 of the shared task. Precision, Recall and F-score over the drug mention class is shown. Top score in each column is shown in bold.

Team	Institution(s)-Country	P	R	F
ClaC	Concordia University, Canada	0.402	0.366	0.383
IIT_KGP	Indian Institute of Technology, India	0.408	0.407	0.408
IRISA	INRIA-IRISA, France	0.434	0.501	0.465
LIGHT	Indian Institute of Technology, India	0.520	0.491	0.505
Techno	University Abou Bekr Belkaid, Algeria	0.327	0.432	0.372
Tub-Oslo	University of Tubingen, Germany University of Oslo, Norway	0.478	0.458	0.468
UChicagoCompLx	University of Chicago, USA	**0.654**	**0.783**	**0.713**
UZH	University of Zurich, Switzerland	0.371	0.437	0.401

Table 2: System performances for each team for task 2 of the shared task. Micro-averaged Precision, Recall and F-score over the definite intake and possible intake classes are shown. Top score in each column is shown in bold.

ticipants were provided with two sets of annotated data, one with 8,181 tweets and the other with 1,665 tweets, where approximately one third of the tweets are labeled positive. Tweets were annotated with binary labels indicating whether the user intends to receive a flu vaccine. The evaluation set consisted of 161 tweets. The evaluation metric for this task was the F-score for the positive class, since the primary intent of this task is to identify if someone has received a flu vaccine.

3 Results

Task 1: Fourteen teams registered to participate in the task and 23 submissions from eleven teams were included in the final evaluations. Table 1 presents the performances of the best systems for each teams having submitted. Team THU_NGN had the best performing system for this task, obtaining a F-score of 0.9182.

Task 2: Eight teams submitted twenty system runs for the final evaluations. Table 2 presents the performances of the best systems for each team in terms of micro-averaged F-score for the intake and possible intake classes. UChicagoCompLx achieved top spot with a micro-averaged F-score of 0.71.

Task 3: Nine teams submitted eighteen system runs for the final evaluations. Table 3 presents the performances of the best systems for each team in terms of ADR class F-score. Team THU_NGN obtained the best F-score of 0.522.

Task 4: Nine teams submitted seventeen system runs for the final evaluations. Table 4 presents the performances of the best systems for each team. Team CARRDS obtained the best F-score of 0.887.

4 Conclusion

The submitted systems employed a wide range of deep learning based classifiers but also feature-based classifiers and few attempts with ensemble learning systems. The system descriptions

Team	Institution(s)-Country	**P**	**R**	**F**
ART	Tata Consultancy Services Limited, India	0.332	0.547	0.413
CIC-NLP	Instituto Politecnico Nacional, Mexico	0.314	0.529	0.394
IIT_KGP	Indian Institute of Technology, India	0.189	0.643	0.292
IRISA	INRIA-IRISA, France	0.378	**0.649**	0.478
Techno	University Abou Bekr Belkaid, Algeria	0.434	0.344	0.383
THU_NGN	Tsinghua University, China	0.442	0.636	**0.522**
Tub-Oslo	University of Tubingen, Germany University of Oslo, Norway	**0.638**	0.317	0.424
UChicagoCompLx	University of Chicago, USA	0.370	0.464	0.411
UZH	University of Zurich, Switzerland	0.455	0.436	0.445

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

Team	Institution(s)-Country	**P**	**R**	**F**
CARRDS	CSIRO-Data61, Australia	**0.918**	0.859	**0.887**
ClaC	Concordia University, Canada	0.700	0.897	0.787
IRISA	INRIA-IRISA, France	0.867	0.833	0.850
IIT_KGP	Indian Institute of Technology, India	0.800	0.769	0.784
LIGHT	Indian Institute of Technology, India	0.824	0.897	0.859
LILU	Technical University of Moldova, Moldova	0.829	0.808	0.818
techno	University Abou Bekr Belkaid, Algeria	0.870	0.859	0.865
Tub-Oslo	University of Tubingen, Germany University of Oslo, Norway	0.840	0.872	0.855
UChicagoCompLx	University of Chicago, USA	0.791	**0.923**	0.852

Table 4: System performances for each team for task 4 of the shared task. Precision, Recall and F-score over the positive class is shown. Top score in each column is shown in bold.

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.

References

Xiaolei Huang, Michael C. Smith, Michael J. Paul, Dmytro Ryzhkov, Sandra C. Quinn, David A. Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Joint Workshop on Health Intelligence*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M. Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. 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. *Journal of the American Medical Informatics Association*, article in press; doi:10.1093/jamia/ocy114.