Overview of the Fourth Social Media Mining for Health (#SMM4H) Shared Task at ACL 2019

Davy Weissenbacher[†], Abeed Sarker[†], Arjun Magge[◇], Ashlynn Daughton[‡], Karen O'Connor[†], Michael Paul[‡], Graciela Gonzalez-Hernandez[†]
[†]DBEI, Perelman School of Medicine, University of Pennsylvania, PA, USA
[◇]Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University, AZ, USA
[‡]Information Science University of Colorado Boulder, CO, USA
{dweissen,abeed,gragon}@pennmedicine.upenn.edu, amaggera@asu.edu
{mpaul,ashlynn.daughton}@colorado.edu

Abstract

The number of users of social media continues to grow, with nearly half of adults worldwide and two-thirds of all American adults using social networking on a regular basis[1]. Advances in automated data processing and NLP present the possibility of utilizing this massive data source for biomedical and public health applications, if researchers address the methodological challenges unique to this media. We present the Social Media Mining for Health Shared Tasks collocated with the ACL at Florence in 2019, which address these challenges for health monitoring and surveillance, utilizing state of the art techniques for processing noisy, real-world, and substantially creative language expressions from social media users. For the fourth execution of this challenge, we proposed four different tasks. Task 1 asked participants to distinguish tweets reporting an adverse drug reaction (ADR) from those that do not. Task 2, a follow-up to Task 1, asked participants to identify the span of text in tweets reporting ADRs. Task 3 is an end-to-end task where the goal was to first detect tweets mentioning an ADR and then map the extracted colloquial mentions of ADRs in the tweets to their corresponding standard concept IDs in the MedDRA vocabulary. Finally, Task 4 asked participants to classify whether a tweet contains a personal mention of one's health, a more general discussion of the health issue, or is an unrelated mention. A total of 34 teams from around the world registered and 19 teams from 12 countries submitted a system run. We summarize here the corpora for this challenge which are freely available at https://competitions.codalab.org/competitions/22521, and present an overview of the methods and the results of the competing systems.

[1]Pew Research Center. Social Media Fact Sheet. 2017. [Online]. Available: http://www.pewinternet.org/fact-sheet/social-media/

1 Introduction

The intent of the #SMM4H shared tasks series is to challenge the community with Natural Language Processing tasks for mining relevant data for health monitoring and surveillance in social media. Such challenges require processing imbalanced, noisy, real-world, and substantially creative language expressions from social media. The competing systems should be able to deal with many linguistic variations and semantic complexities in the various ways people express medication-related concepts and outcomes. It has been shown in past research (Liu et al., 2011; Giuseppe et al., 2017) that automated systems frequently under-perform when exposed to social media text because of the presence of novel/creative phrases, misspellings and frequent use of idiomatic, ambiguous and sarcastic expressions. The tasks act as a discovery and verification process of what approaches work best for social media data.

As in previous years, our tasks focused on mining health information from Twitter. This year we challenged the community with two different problems. The first problem focuses on performing pharmacovigilance from social media data. It is now well understood that social media data may contain reports of adverse drug reactions (ADRs) and these reports may complement traditional adverse event reporting systems, such as the FDA adverse event reporting system (FAERS). However, automatically curating reports from adverse reactions from Twitter requires the application of a series of NLP methods in an end-to-end pipeline (Sarker et al., 2015). The first three tasks of this year's challenge represent three key NLP problems in a social media based pharmacovigilance pipeline — (i) automatic classification of ADRs, (ii) extraction of spans of ADRs and (iii) normal-

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 21–30
Florence, Italy, August 2, 2019.

ization of the extracted ADRs to standardized IDs.

The second problem explores the generalizability of predictive models. In health research using social media, it is often necessary for researchers to build individual classifiers to identify health mentions of a particular disease in a particular context. Classification models that can generalize to different health contexts would be greatly beneficial to researchers in these fields (e.g., (Payam and Eugene, 2018)), as this would allow researchers to more easily apply existing tools and resources to new problems. Motivated by these ideas, Task 4 was testing tweet classification methods across diverse health contexts, so the test data included a very different health context than the training data. This setting measures the ability of tweet classifiers to generalize across health contexts.

The fourth iteration of our series follows the same organization as previous iterations. We collected posts from Twitter, annotated the data for the four tasks proposed and released the posts to the registered teams. This year, we conducted the evaluation of all participating systems using Codalab, an open source platform facilitating data science competitions. The performances of the systems were compared on a blind evaluations sets for each task.

All teams registered were allowed to participate to one or multiple tasks. We provided the participants with two sets of data for each task, a training and a test set. Participants had a period of six weeks, from March 5^{th} to April 15^{th}, for training their systems on our training sets, and 4 days, from the 16^{th} to 20^{th} of April, for calibrating their systems on our test sets and submitting their predictions. In total 34 teams registered and 19 teams submitted at least one run (each team was allowed to submit, at most, three runs per task). In detail, we received 43 runs for task 1, 24 for task 2, 10 for task 3 and 15 for task 4. We briefly describe each task and their data in section 2, before discussing the results obtained in section 3.

2 Task Descriptions

2.1 Tasks

Task 1: Automatic classification of tweets mentioning an ADR. This is a binary classification task for which systems are required to predict if a tweet mentions an ADR or not. In an end-to-end social media based pharmacovigilance pipeline, such a system is needed after data collection to filter out the large volume of medication-related chatter that is not a mention of an ADR. This task is a rerun of the popular classification task organized in past years.

Task 2: Automatic extraction of ADR mentions from tweets. This is a named entity recognition (NER) task that typically follows the ADR classification step (Task 1) in an ADR extraction pipeline. Given a set of tweets containing drug mentions and potentially containing ADRs, the objective was to determine the span of the ADR mention, if any. ADRs are rare events making ADR classification a challenging task with an F1-score in the vicinity of 0.5 (based on previous shared task results (Weissenbacher et al., 2018)) for the ADR class. The dataset for the ADR extraction task contains tweets that are both positive and negative for the presence of ADRs. This allowed participants to choose to train their systems on either the set of tweets containing ADRs or include tweets that were negative for the presence of ADRs.

Task 3: Automatic extraction of ADR mentions and normalization of extracted ADRs to MedDRA preferred term identifiers. This is an extension of Task 2 consisting of the combination of NER and entity normalization tasks: a named entity resolution task. In this task, given the same set of tweets as in Task 2, the objective was to extract the span of an ADR mention and to normalize it to MedDRA identifiers [2]. MedDRA (Medical Dictionary for Regulatory Activities), which is the standard nomenclature for monitoring medical products, and includes diseases, disorders, signs, symptoms, adverse events or adverse drug reactions. For the normalization task, MedDRA version 21.1 was used, containing 79,507 lower level terms (LLTs) and 23,389 respective preferred terms (PTs).

Task 4: Automatic classification of personal mentions of health. In this binary classification task, the systems were required to distinguish tweets of personal health status or opinions across different health domains. The proposed task was intended to provide a baseline understanding of the ability to identify personal health mentions in a generalized context.

[2] https://www.meddra.org/ Accessed: 05/13/2019.

2.2 Data

All corpora were composed of public tweets downloaded using the official streaming API provided by Twitter and made available to the participants in accordance with Twitter's data use policy. This study received an exempt determination by the Institutional Review Board of the University of Pennsylvania.

Task 1. For training, participants were provided with all the tweets from the #SMM4H 2017 shared tasks (Sarker et al., 2018), which are publicly available at: https://data.mendeley.com/datasets/rxwfb3tysd/2. A total of 25,678 tweets were made available for training. The test set consisted of 4575 tweets with 626 (13.7%) tweets representing ADRs. The evaluation metric for this task was micro-averaged F1-score for the ADR class.

Task 2. Participants of Task 2 were provided with a training set containing 2276 tweets which mentioned at least one drug name. The dataset contained 1300 tweets that were positive for the presence of ADRs and 976 tweets that were negative. Participants were allowed to include additional negative instances from Task 1 for training purposes. Positive tweets were annotated with the start and end indices of the ADRs and the corresponding span text in the tweets. The evaluation set contained 1573 tweets, 785 and 788 tweets were positive and negative for the presence of ADRs respectively. The participants were asked to submit outputs from their systems that contained the predicted start and end indices of ADRs. The participants' submissions were evaluated using standard strict and overlapping F1-scores for extracted ADRs. Under strict mode of evaluation, ADR spans were considered correct only if both start and end indices matched with the indices in our gold standard annotations. Under overlapping mode of evaluation, ADR spans were considered correct only if spans in predicted annotations overlapped with our gold standard annotations.

Task 3. Participants were provided with the same training and evaluation datasets as in Task 2. However, the datasets contained additional columns for the MedDRA annotated LLT and PT identifiers for each ADR mention. In total, of the 79,507 LLT and 23,389 PT identifiers available in MedDRA, the training set of 2276 tweets and 1832 annotated ADRs contained 490 unique LLT identifiers and 327 unique PT identifiers. The evaluation set contained 112 PT identifiers that were not present as part of the training set. The participants were asked to submit outputs containing the predicted start and end indices of ADRs and respective PT identifiers. Although the training dataset contained annotations at the LLT level, the performance was only evaluated at the higher PT level. The participants' submissions were evaluated using standard strict and overlapping F-scores for extracted ADRs and respective MedDRA identifiers. Under strict mode of evaluation, ADR spans were considered correct only if both start and end indices matched along with matching MedDRA PT identifiers. Under overlapping mode of evaluation, ADR spans were considered correct only if spans in predicted ADRs overlapped with gold standard ADR spans in addition to matching MedDRA PT identifiers.

Task 4 Data. Participants were provided training data from one disease domain, influenza, across two contexts, being sick and getting vaccinated, both annotated for personal mentions: the user is personally sick or the user has been personally vaccinated. Test data included new tweets of personal health mentions about influenza and tweets from an additional disease domain, Zika virus, with two different contexts, the user is changing their travel plans in response to Zika concerns, or the user is minimizing potential mosquito exposure due to Zika concerns.

2.3 Annotation and Inter-Annotator Agreements

Two annotators with biomedical education and both experienced in Social Media research tasks manually annotated the corpora for tasks 1, 2 and 3. Our annotators independently dual-annotated each test sets to insure the quality of our annotations. Disagreement were resolved after an adjudication phase between our two annotators. On task 1, the classification task, the inter annotator-agreement (IAA) was high with a Cohens Kappa = 0.82. On task 2, the information extraction task, IAAs were good with and an F1-score of 0.73 for strict agreement, and 0.85 for overlapping agreement[3]. On task 3, our annotators double annotated

[3]Since task 2 is a named-entity recognition task, we followed the recommendations of (Hripcsak and Rothschild, 2005) and used precision and recall metrics to estimate the inter-annotator rate.

535 of the extracted ADR terms and normalized them to MedDRA lower lever terms (LLT). They achieved an agreement accuracy of 82.6%. After converting the LLT to their corresponding preferred term (PT) in MedDRA, which is the coding the task was scored against, accuracy improved to 87.7%[4].

The annotation process followed for task 4 was slightly different due to the nature of the task. We obtained the two datasets of our training set, focusing on flu vaccination and flu infection, from (Huang et al., 2017) and (Lamb et al., 2013) respectively. Huang et al. (Huang et al., 2017) used mechanical turk to crowdsource labels (Fleiss' kappa = 0.793). Lamb et al. (Lamb et al., 2013) did not report their labeling procedure or annotator agreement metrics, but do report annotation guidelines[5]. A few of the tweets released by Lamb et al. appeared to be mislabeled and were corrected in accordance with the annotation guidelines defined by the authors. We obtained the test data for task 4 by compiling three datasets. For the dataset related to travel changes due to Zika concerns, we selected a subset of data already available from (Daughton and Paul, 2019). Initial labeling of these tweets was performed by two annotators with a public health background (Cohen's kappa = 0.66). We reuse the original annotations for this dataset without changes. For the mosquito exposure dataset, tweets were labeled by one annotator with public health knowledge and experienced with social media, and then verified by a second annotator with similar experience. The additional set of data on personal exposure to Influenza were obtained from a separate group, who used an independent labeling procedure.

3 Results

The challenge received a solid response with 19 teams from 12 countries (7 from North America, 1 from South America, 6 from Asia and 5 from Europe) submitting 92 runs in total in one or more tasks. We present an overview of all architectures competing in the different tasks in Table 1, 2, 3, 4. We also list in these tables the external resources competitors integrated for improving the pre-training of their systems or for embedding high-level features to help decision-making.

The overview of all architectures is interesting in two ways. First, this challenge confirms the tendency of the community to abandon traditional Machine Learning systems based on handcrafted features for deep learning architectures capable of discovering the features relevant for the task at hand from pre-trained embeddings. During the challenge, when participants implemented traditional systems, such as SVM or CRF, they used such systems as baselines and, observing significant differences of performances with systems based on deep learning on their validation sets, most of them did not submit their predictions as official runs. Second, while last year convolutional or recurrent neural networks "fed" with pre-trained word embeddings learned on local windows of words (*e.g.* word2vec, GloVe) were the most popular architectures, this year we can see a clear dominance of neural architectures using word embeddings pre-trained with the Bidirectional Encoder Representations from Transformers (BERT) proposed by (Devlin et al., 2018), or fine-tuning these words embeddings on our training corpora. BERT allows to compute words embeddings based on the full context of sentences and not only on local windows.

A notable result from task 1-3 is that, despite an improvement in performances for the detection of ADRs, their resolution remains challenging and will require further research. The participants largely adopted contextual word-embeddings during this challenge, a choice rewarded by new records in performances during the task 1, the only task reran from last years. The performances increased from .522 F1-score (.442 P, .636 R) (Weissenbacher et al., 2018) to .646 F1-score (0.608 P, 0.689 R) for the best systems of each years. However, with a strict matching F1-score of .432 (.362 P, .535 R) for the best system, the performances obtained in task 3 for ADRs resolution are still low and human inspection is still required to make use of the data extracted automatically. As shown by the best score of .887 Accuracy obtained on the ADR normalization in task 3 ran during #SMM4H in 2017 (Sarker et al., 2018)[6], once ADRs are extracted, the normalization of the ADRs can be per-

[4]We measured agreement using accuracy instead of Cohens Kappa because, with greater than 70,000 LLTs for the annotators to choose from, agreement due to chance is expected to be small.

[5]We used the awareness vs. infection labels as defined in (Lamb et al., 2013).

[6]Organizers of the task 3 ran during #SMM4H 2017 provided participants with manually curated expressions referring to ADRs and participants had to map them to their corresponding preferred terms in MeDRA.

formed with a good reliability. However errors are made during all steps of the resolution — detection, extraction, normalization — and their overall accumulation render current automatic systems inefficient. Note that bulk of the errors are made during the extraction of the ADRs, as shown by the low strict F1-score of the best system in task 2, .464 F1-score (.389P, .576 R).

For task 4, we were especially interested in the generalizability of first person health classifiers to a domain separate from that of the training data. We find that, on average, teams do reasonably well across the full test dataset (average F1-score: 0.70, range: 0.41-0.87). Unsurprisingly, classifiers tended to do better on a test set in the same domain as the training dataset (context 1, average F1-score: 0.82) and more modestly on the Zika travel and mosquito datasets (average F1-score: 0.40 and 0.52, respectively). Interestingly, in all contexts, precision was higher than recall. We note that both the training and the testing data were limited in quantity, and that classifiers would likely improve with more data. However, in general, it is encouraging that classifiers trained in one health domain can be applied to separate health domains.

4 Conclusion

In this paper we presented an overview of the results of #SMM4H 2019 which focuses on a) the resolution of adverse drug reaction (ADR) mentioned in Twitter and b) the distinction between tweets reporting personal health status form opinions across different health domains. With a total of 92 runs submitted by 19 teams, the challenge was well attended. The participants, in large part, opted for neural architectures and integrated pretrained word-embedding sensitive to their contexts based on the recent Bidirectional Encoder Representations from Transformers. Such architectures were the most efficient on our four tasks. Results on tasks 1-3 show that, despite a continuous improvement of performances in the detection of tweets mentioning ADRs over the past years, their end-to-end resolution still remain a major challenge for the community and an opportunity for further research. Results of task 4 were more encouraging, with systems able to generalized their predictions over domains not present in their training data.

References

Ashlynn R. Daughton and Michael J. Paul. 2019. Identifying protective health behaviors on twitter: Observational study of travel advisories and zika virus. *Journal of Medical Internet Research. In Press.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rizzo Giuseppe, Pereira Bianca, Varga Andrea, van Erp Marieke, and Elizabeth Cano Basave Amparo. 2017. Lessons learnt from the named entity recognition and linking (neel) challenge series. *Semantic Web Journal*, 8(5):667–700.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

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 (W3PHIAI)*.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367. Association for Computational Linguistics.

Zulfat Miftahutdinov, Elena Tutubalina, and Alexander Tropsha. 2017. Identifying disease-related expressions in reviews using conditional random fields. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog*, volume 1, pages 155–167.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez-Hernandez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mention using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Karisani Payam and Agichtein Eugene. 2018. Did you really just have a heart attack? towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 137–146.

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 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. *J Am Med Inform Assoc*, 25(10):1274–1283.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202 – 212.

Abeed Sarker and Graciela Gonzalez. 2017. A corpus for mining drug-related knowledge from twitter chatter: Language models and their utilities. *Data in Brief*, 10:122–131.

Abeed Sarker and Graciela Gonzalez-Hernandez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Davy Weissenbacher, Abeed Sarker, Michael J Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *in Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 13–16.

Abbreviations

FF: Feedforward
CNN: Convolutional Neural Network
BiLSTM: Bidirectional Long Short-Term Memory
SVM: Support Vector Machine
CRF: Conditional Random Field
POS: Part-Of-Speech
RNN: Recurrent Neural Network

Rank	Team	System details
1	ICRC	*Architecture:* BERT + FF + Softmax *Details:* lexicon features (pairs of drug-ADR) *Resources:* SIDER
2	UZH	*Architecture:* ensemble of BERT & C_CNN + W_BiLSTM (+ CRF) *Details:* multi-task-learning *Resources:* CADEC corpus
3	MIDAS@IIITD	*Architecture:* 1. BERT 2. ULMFit 3. W_BiLSTM *Details:* BERT + GloVe + Flair *Resources:* additional corpus (Sarker and Gonzalez-Hernandez, 2015)
4	KFU NLP	*Architecture:* BERT + logistic regression *Details:* BioBERT
5	CLaC	*Architecture:* Bert + W_BiLSTM + attention + softmax + SVM *Details:* BERT, Word2Vec, Glove, embedded features *Resources:* POS, modality, ADR list
6	THU_NGN	*Architecture:* C_CNN + W_BiLSTM + features + Multi-Head attention + Softmax *Details:* Word2Vec, POS, ELMo *Resources:* sentiment Lexicon, SIDER, CADEC
7	BigODM	*Architecture:* ensemble of SVMs *Resources:* Word Embeddings
8	UMich-NLP4Health	*Architecture:* 1. W_BiLSTM + attention + softmax; 2. W_CNN + BiLSTM + softmax; 3. SVM *Details:* GloVe, POS, case *Resources:* Metamap, cTAKES, CIDER
9	TMRLeiden	*Architecture:* ULMfit *Details:* Flair + Glove + Bert; transfer learning *Resources:* external corpus (Sarker and Gonzalez, 2017)
10	CIC-NLP	*Architecture:* C_BiLSTM + W_FF + LSTM + FF *Details:* GloVe + BERT
12	SINAI	*Architecture:* 1. SVM 2. CNN + Softmax *Details:* GloVe *Resources:* MetaMap
13	nlp-uned	*Architecture:* W_BiLSTM + Sigmoid *Details:* GloVe
14	ASU BioNLP	*Architecture:* 1. Lexicon; 2. BioBert *Details:* Lexicon learned with Logistic regression model
15	Klick Health	*Architecture:* ELMo + FF + Softmax *Details:* Lexicons *Resources:* MedDRA, Consumer Health Vocabulary, (Nikfarjam et al., 2015)
16	GMU	*Architecture:* encoder-decoder (W_biLSTM + attention) *Details:* Glove *Resources:* #SMM4H 2017-2018, UMLS

Table 1: Task 1. System and resource descriptions for ADR mentions detection in tweets[7].

[8] We use C_BiLSMT and C_CNN to denote bidirectonal LSTMs or CNNs encoding sequences of characters, W_BiLSTM and W_FF to denote bidirectional LSTMs or Feed Forward encoders of word embeddings.

Rank	Team	System details
1	KFU NLP	*Architecture:* ensemble of BioBERT + CRF *Details:* BioBERT *Resources:* external dictionaries (Miftahutdinov et al., 2017); CADEC, PsyTAR, TwADR-L corpora; #SMM4H 2017
2	THU_NGN	*Architecture:* C_CNN + W_BiLSTM + features + Multi-Head self-attention + CRF *Details:* Word2Vec, POS, ELMo *Resources:* sentiment Lexicon, SIDER, CADEC
3	MIDAS@IIITD	*Architecture:* W_BiLSTM + CRF *Details:* BERT + GloVe + Flair
4	TMRLeiden	*Architecture:* BERT + Flair *Details:* Flair + Glove + Bert; transfer learning
5	ICRC	*Architecture:* BERT + CRF *Resources:* SIDER
6	GMU	*Architecture:* C_biLSTM + W_biLSTM + CRF *Details:* Glove *Resources:* #SMM4H 2017-2018, UMLS
7	HealthNLP	*Architecture:* W_BiLSTM + CRF *Details:* Word2vec, BERT, ELMo, POS *Resources:* external dictionaries
8	SINAI	*Architecture:* CRF *Details:* GloVe *Resources:* MetaMap
9		*Architecture:* BiLSTM + CRF *Details:* Word2Vec *Resources:* MIMIC-III
10	Klick Health	*Architecture:* Similarity *Details:* Lexicons *Resources:* MedDRA, Consumer Health Vocabulary, (Nikfarjam et al., 2015)

Table 2: Task 2. System and resource descriptions for ADR mentions extraction in tweets

Rank	Team	System details
1	KFU NLP	*Architecture:* BioBERT + softmax
2	myTomorrows-TUDelft	*Architecture:* ensemble RNN & Few-Shot Learning *Details:* Word2Vec *Resources:* MedDRA, Consumer Health Vocabulary, UMLS
3	TMRLeiden	*Architecture:* BERT + Flair + RNN *Details:* Flair + Glove + Bert; transfer learning *Resources:* Consumer Health Vocabulary
4	GMU	*Architecture:* encoder-decoder (W_biLSTM + attention) *Details:* Glove *Resources:* #SMM4H 2017-2018, UMLS

Table 3: Task 3. System and resource descriptions for ADR mentions resolution in tweets.

Rank	Team	System details
1	UZH	*Architecture:* ensemble BERT *Resources:* CADEC corpus
2	ASU1	*Architecture:* BioBERT + FF *Resources:* Word2vec, manually compiled list, ConceptNet
4	MIDAS@IIITD	*Architecture:* BERT; W_BiLSTM *Details:* BERT + GloVe + Flair
5	TMRLeiden	*Architecture:* ULMfit *Details:* Flair + Glove + Bert; transfer learning *Resources:* external corpus (Payam and Eugene, 2018)
6	CLaC	*Architecture:* Bert + W_BiLSTM + attention + softmax + SVM *Details:* BERT, Word2Vec, Glove, embedded features *Resources:* POS, modality, ADR list

Table 4: Task 4. System and resource descriptions for detection of personal mentions of health in tweets.

Team	F1	P	R
ICRC	**0.6457**	0.6079	0.6885
UZH	0.6048	0.6478	0.5671
MIDAS@IIITD	0.5988	0.6647	0.5447
KFU NLP	0.5738	**0.6914**	0.4904
CLaC	0.5738	0.5427	0.6086
THU_NGN	0.5718	0.4667	**0.738**
BigODM	0.5514	0.4762	0.655
UMich-NLP4Health	0.5369	0.5654	0.5112
TMRLeiden	0.5327	0.6419	0.4553
CIC-NLP	0.5209	0.6203	0.4489
UChicagoCompLx	0.4993	0.4574	0.5495
SINAI	0.4969	0.5517	0.4521
nlp-uned	0.4723	0.5244	0.4297
ASU BioNLP	0.4317	0.3223	0.6534
Klick Health	0.4099	0.5824	0.3163
GMU	0.3587	0.4526	0.2971

Table 5: System performances for each team for task 1 of the shared task. F1-score, Precision and Recall over the ADR class are shown. Top scores in each column are shown in bold.

	Relaxed			Strict		
Team	**F1**	**P**	**R**	**F1**	**P**	**R**
KFU NLP	**0.658**	0.554	**0.81**	**0.464**	0.389	**0.576**
THU_NGN	0.653	0.614	0.697	0.356	0.328	0.388
MIDAS@IIITD	0.641	0.537	0.793	0.328	0.274	0.409
TMRLeiden	0.625	0.555	0.715	0.431	0.381	0.495
ICRC	0.614	0.538	0.716	0.407	0.357	0.474
GMU	0.597	0.596	0.599	0.407	0.406	0.407
HealthNLP	0.574	**0.632**	0.527	0.336	0.37	0.307
SINAI	0.542	0.612	0.486	0.36	**0.408**	0.322
ASU BioNLP	0.535	0.415	0.753	0.269	0.206	0.39
Klick Health	0.396	0.416	0.378	0.194	0.206	0.184

Table 6: System performances for each team for task 2 of the shared task. (Strict/Relaxed) F1-score, Precision and Recall over the ADR mentions are shown. Top scores in each column are shown in bold.

	Relaxed			Strict		
Team	**F1**	**P**	**R**	**F1**	**P**	**R**
KFU NLP	**0.432**	0.362	**0.535**	**0.344**	0.288	**0.427**
myTomorrows-TUDelft	0.345	0.336	0.355	0.244	0.237	0.252
TMRLeiden	0.312	**0.37**	0.27	0.25	**0.296**	0.216
GMU	0.208	0.221	0.196	0.109	0.116	0.102

Table 7: System performances for each team for task 3 of the shared task. (Strict/Relaxed) F1-score, Precision and Recall over the ADR resolution are shown. Top scores in each column are shown in bold.

Team	Acc	F1	P	R
Health concerns in all contexts				
UZH	**0.8772**	**0.8727**	0.8392	**0.9091**
ASU1	0.8456	0.8036	**0.9783**	0.6818
UChicagoCompLx	0.8316	0.7913	0.9286	0.6894
MIDAS@IIITD	0.8211	0.783	0.8932	0.697
TMRLeiden	0.793	0.7256	0.9398	0.5909
CLaC	0.6386	0.4607	0.7458	0.3333
Health concerns in Context 1: Flu virus (infection/vaccination)				
UZH	0.9438	0.9474	0.9101	0.9878
UChicagoCompLx	0.925	0.9231	0.973	0.878
ASU1	0.925	0.9221	0.9861	0.8659
MIDAS@IIITD	0.8875	0.88	0.9706	0.8049
TMRLeiden	0.8625	0.8493	0.9688	0.7561
CLaC	0.6625	0.5645	0.8333	0.4268
Health concerns in Context 2: Zika virus, travel plans changes				
UZH	0.7536	0.7385	0.7059	0.7742
MIDAS@IIITD	0.6667	0.5818	0.6667	0.5161
ASU1	0.6957	0.5116	0.9167	0.3548
UChicagoCompLx	0.6377	0.4681	0.6875	0.3548
TMRLeiden	0.6377	0.4186	0.75	0.2903
CLaC	0.5362	0.2	0.4444	0.129
Health concerns in Context 3: Zika virus, reducing mosquito exposure				
UZH	0.8393	0.7692	0.75	0.7895
MIDAS@IIITD	0.8214	0.6667	0.9091	0.5263
ASU1	0.8036	0.5926	1.0	0.4211
UChicagoCompLx	0.8036	0.5926	1.0	0.4211
TMRLeiden	0.7857	0.5385	1.0	0.3684
CLaC	0.6964	0.3704	0.625	0.2632

Table 8: System performances for each team for task 4 of the shared task. Accuracy, F1-score, Precision and Recall over the personal mentions are shown. Top scores in each column are shown in bold.