Towards text processing pipelines to identify adverse drug events-related tweets: University of Michigan @ SMM4H 2019 Task 1

V.G.Vinod Vydiswaran,[1,2*] **Grace Ganzel,**[2] **Bryan Romas,**[2] **Deahan Yu,**[2]
Amy M. Austin,[2] **Neha Bhomia,**[2] **Socheatha Chan,**[2] **Stephanie V. Hall,**[1] **Van Le,**[2]
Aaron Miller,[2] **Olawunmi Oduyebo,**[2] **Aulia Song,**[2] **Radhika Sondhi,**[2] **Danny Teng,**[2]
Hao Tseng,[2] **Kim Vuong,**[2] **Stephanie Zimmerman**[2]
[1]Department of Learning Health Sciences, [2]School of Information, University of Michigan
*Corresponding author: vgvinodv@umich.edu

Abstract

We participated in Task 1 of the Social Media Mining for Health Applications (SMM4H) 2019 Shared Tasks on detecting mentions of adverse drug events (ADEs) in tweets. Our approach relied on a text processing pipeline for tweets, and training traditional machine learning and deep learning models. Our submitted runs performed above average for the task.

1 Introduction

A growing number of users produce and share information on the internet, including health information. As of 2017, the number of social media users increased by approximately one million per day, with approximately half of adults worldwide using some form of social media (Kemp, 2017). According to (Domo, 2018), 473,400 tweets were sent every minute in 2018.

The discussion of health-related information on social media is becoming increasingly common, which can be utilized by researchers in a multitude of ways, including pharmacovigilance and public health surveillance (Nikfarjam et al., 2015). Numerous works have utilized tweets to analyze public health concerns, with many focusing specifically on the identification of adverse drug reactions (ADRs) (O'Connor et al., 2014). Tweets can contain important information, including the mention of specific medications, indications for use, and side effects. Additional information can also be obtained, such as time of the tweet, location, and user characteristics (Paul and Dredze, 2011).

The Social Media Mining for Health Applications (SMM4H) Shared Tasks were organized to provide labeled social media data sets for natural language processing (NLP) researchers to study challenges in health monitoring and surveillance (Weissenbacher et al., 2019). Task 1 focused on classification of adverse drug event (ADE) mentions in tweets, where the participating systems were expected to distinguish tweets reporting an ADE from those that do not, taking into account subtle linguistic variations between adverse effects and indications, such as the reason to use the medication.

2 Data cleaning and pre-processing

Our approach was to develop a text processing pipeline to clean and process tweets and identify tweets that mention ADEs. We ran the following pre-processing steps:

(a) Removing UTF-8 characters: All UTF-8 characters in the tweets were removed or replaced with relevant tags. For example, a pill emoji was replaced with the tag '⟨pill⟩', and a dizzy-faced emoji was replaced with the tag '⟨dizzy⟩'.

(b) Running Ekphrasis: After all UTF-8 characters were removed, the Ekphrasis text processing tool (Baziotis et al., 2017) was run with the following minor modifications. First, because the tool was unable to unpack contractions that appeared in uppercase text, regular expressions were written to capture all uppercase tokens for manual verification and tagging. After tagging, the tweets were converted to lowercase, allowing them to be fully processed by the unpacking feature. New contractions were added to the Ekphrasis unpacking routine based on a manual review of Ekphrasis output, when applied to the challenge data set.

(c) NLTK TweetTokenizer and Lemmatizer: NLTK TweetTokenizer was run to further process the tweets, and the outputted tweets were then lemmatized.

(d) MetaMap: Each tweet was run through MetaMap, and concept and semantic types identified within the text with a MetaMap score above 800 were extracted as features.

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 107–109
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

(e) cTAKES: Tweets were run through cTAKES to identify concepts from the Systematized Nomenclature of Medicine (SNOMED). The identified SNOMED codes were added as features.

(f) Pattern-based features: Additional features were generated based on pattern-matching rules using regular expressions.

3 Word representation for neural models

We trained two variations of neural network models — a bidirectional LSTM (Graves and Schmidhuber, 2005) model, and a bidirectional LSTM model with a convolutional neural network (CNN) layer (Kim, 2014). We also compared the performance of both models using pre-trained GloVe word embedding (Pennington et al., 2014) and using pre-trained Word2vec Twitter word embedding (Godin et al., 2015). To evaluate the performance of the models, we randomly split the training set in a 80-20 ratio while maintaining the original class proportions. The four models were trained on 80% of the provided training data and tested on the remaining 20% (validation data set). Of the four models, the best model based on validation accuracy was chosen as the final model, which was the bidirectional LSTM model using GloVe word embedding.

3.1 Features

To generate the input tweet representation for deep learning models, we undertook the following additional steps:

(a) Part-of-speech tag embedding: To create a part-of-speech (POS) embedding, we used NLTK to first extract POS labels for each word. We then converted each tweet into a sequence of POS tags according to the token order and created the POS tag embedding.

(b) First-character embedding: Similar to the part-of-speech tag embedding, we extracted the first character of each token in a tweet and generated four binary features depending on whether the first character was an uppercase letter, a lowercase letter, an integer, or a symbol / special character.

(c) Medical dictionary: Finally, we obtained a MedDRA dictionary from Side Effect Resource, SIDER (Kuhn et al., 2016, 2010) and used it to create a one-hot vector representation for words listed in SIDER, in addition to word embedding.

4 Description of runs

Once the pre-processing and input representation were finalized, we trained the following three models corresponding to the three submitted runs:

Run 1: As a baseline for our models, we trained a linear kernel support vector machine classifier with balanced class weights. The model was trained over unigram features generated from the lowercased tweet text. The individual feature weights were computed using their inverse document frequency over the training data set. The classifier was built using scikit-learn.

Run 2: For the second run, we ran all tweets through the pre-processing pipeline described in Sec. 2. The tweet text was cleaned using the modified Ekphrasis tool, features from MetaMap and cTAKES were added, and the text was tokenized. Unigram and bigram features were instantiated and were weighted by the inverse document frequency in the training set. A linear kernel support vector machine classifier was trained with a balanced class weight configuration.

Run 3: For the third run, we used a bidirectional LSTM with categorical cross entropy loss function with RMSprop optimizer. We set the model dropout layer probability to 0.2 in order to avoid overfitting. Following (Vaswani et al., 2017), we added an attention layer. Our output layer for the classification task was a dense layer followed by the softmax function. For the input representation, we employed a concatenation of the pre-trained GloVe word embedding and the first character embedding. We padded each tweet to 29 tokens, which is the sum of the average tweet length ($\ell = 16$) and two standard deviations of the tweet lengths ($\sigma = 6.5$) in the 80% data set. We set it this way because the maximum length from the 80% of the data was too long ($\ell = 130$ tokens) to use and the average length was too short to cover substantial amount of tweets. The model was trained on the 100% of the provided data (both training and validation sets) and run for 100 epochs.

5 Results

In all, 16 teams participated in Task 1 for a total of 43 runs. Table 1 summarizes the performance of our three runs and the average over all runs submitted to the task. Runs 1 and 2 were better than the average performance over recall and F1 mea-

Run ID	Pred pos (%)	Prec	Rec	F1
Run 1	865 (18.9%)	0.452	**0.625**	0.525
Run 2	566 (12.4%)	**0.565**	0.511	**0.537**
Run 3	492 (10.8%)	0.555	0.436	0.488
Avg. task performance		0.535	0.505	0.502

Table 1: Performance of the submitted runs in terms of count (and percentage) of predicted positive instances, precision, recall, and F1 over the test set (n = 4,575).

sures, while runs 2 and 3 were better than the average run on precision. Run 2 was the best among the three submitted runs. It identified 566 (12.4%) tweets as positive with a precision of 0.565, recall of 0.511, and F1 measure of 0.537. All these measures were better than the average measures among runs submitted for Task 1.

6 Conclusion

Our approach for participating in the 2019 SMM4H Shared Task 1 was to develop a text processing pipeline for tweets, focusing on preprocessing, feature weighting, and training traditional feature-based and deep learning models. Our runs performed above the average shared task performance, and the best run achieved an F1 measure of 0.537. Additional runs are planned to further analyze the performance of deep learning models on this task.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. ACL.

Domo. 2018. Data never sleeps 6.0. Retrieved from https://www.domo.com/learn/data-never-sleeps-6.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia Lab @ ACL WNUT NER Shared Task: Named entity recognition for Twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153, Beijing, China. ACL.

Alex Graves and Jurgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, volume 4, pages 2047–2052.

Simon Kemp. 2017. The global state of the Internet in April 2017. Retrieved from https://thenextweb.com/contributors/2017/04/11/current-global-state-internet/.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the Association for Computational Linguistics*, pages 1746–1751.

Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6:343.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(Database issue):D1075–D1079.

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

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *Proceedings of the AMIA Annual Symposium*, pages 924–933.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 265–272.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth Social Media Mining for Health (SMM4H) shared task at ACL 2019. In *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task*.