Identification of Adverse Drug Reaction Mentions in Tweets – SMM4H Shared Task 2019

Samarth Rawal Arizona State University samrawal@asu.edu

Siddharth Rawal Arizona State University sidrawal@asu.edu

Saadat Anwar Arizona State University sanwar@asu.edu

Chitta Baral Arizona State University chitta@asu.edu

Abstract

Analyzing social media posts can offer insights into a wide range of topics that are commonly discussed online, providing valuable information for studying various health-related phenomena reported online. The outcome of this work can offer insights into pharmacovigilance research to monitor the adverse effects of medications. This research specifically looks into mentions of adverse drug reactions (ADRs) in Twitter data through the Social Media Mining for Health Applications (SMM4H) Shared Task 2019. Adverse drug reactions are undesired harmful effects which can arise from medication or other methods of treatment. The goal of this research is to build accurate models using natural language processing techniques to detect reports of adverse drug reactions in Twitter data and extract these words or phrases.

1 Introduction

On average, one in a thousand messages from public Twitter data is health-related (Sadilek et al., 2012). These health-related Twitter posts can be used to monitor and analyze various health-related phenomena such as drug use and side effects resulting from medication. The purpose of this work was to develop a model to accurately analyze mentions of adverse drug reactions (ADRs) in Twitter posts. To achieve this task, natural language processing techniques were used to predict whether each Tweet from a given set of Tweets contains a mention of an ADR and extract any mentions of ADRs. The results of this project can be useful for research done in the field of pharmacovigilance, which is the monitoring of drug effects with the intention of finding and preventing adverse effects. This work was conducted as part of the Social Media Mining for Health (SMM4H) challenge hosted by the Health Language Processing (HLP) Lab at the University of Pennsylvania. The predictions of the models developed for this project were evaluated against test data and given F-scores as well as scores of accuracy, precision, and recall based on the degree to which they were able to accomplish the goals of each task.

2 Methods

2.1 Subtask 1

For Subtask 1, a lexicon-based approach was followed. To identify important keywords – keywords whose presence or absence in a Tweet can serve as valuable, reliable indicators of whether the Tweet contains a reference to an Adverse Drug Reaction or not – a methodology adapted from the Internal + External Lexicon Selection technique (Rawal et al., 2019), a technique that has yielded successful results in previous similar classification tasks, was used. First, uni- and bi-grams were extracted from the training dataset. The presence or absence of each of these n-grams were then used as binary features in a logistic regression model. To estimate the performance of the model using metrics that were to be used for evaluation, such as precision, recall, and F1-score, the model was trained via 10-fold cross-validation of the training set. Finally, the coefficients associated with each keyword were examined. There were 166,466 total features obtained through the aforementioned technique. Through this process, the top 700 absolute-valued coefficients were hypothesized to be the most significant keywords and stored. This number of top keywords to keep was a hyperparameter that was experimentally determined through model performance over 10-fold cross-validation of the training set. This list of significant keywords was then manually pared down to exclude any intuitively irrelevant terms (such as stop words); the presence or absence of these remaining keywords were used as binary features for our final logistic regression model. Other mod-

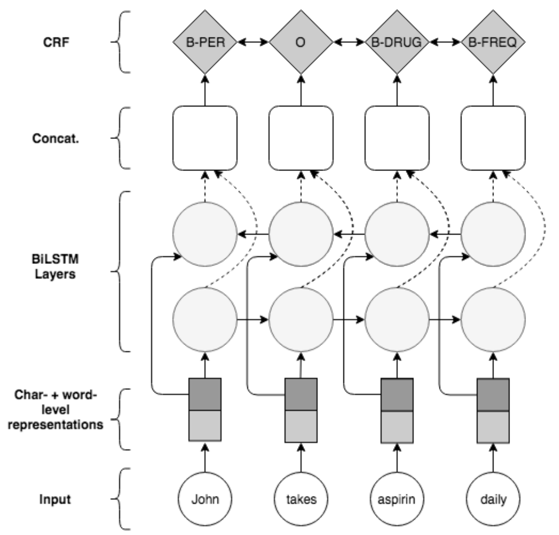
Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 136–137
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

els were also tested during training, such as a BioBERT (Lee et al., 2019) model that was fine-tuned using the provided training data. Although the BioBERT model showed promising results, it was not implemented into the final submission due to time constraints.

2.2 Subtask 2

For Subtask 2, a deep learning approach was taken. Specifically, a Bidirectional Long Short-Term Memory (BiLSTM) coupled with a Conditional Random Field (CRF) layer neural network architecture was used to perform Named Entity Recognition to identify the Adverse Drug Reaction mentions. This architecture has been empirically shown to perform well at Named Entity Recognition (NER) tasks (Lample et al., 2016). To represent input words, the Embedding layer weights of the model was pre-initialized with values obtained from a word2vec model that was trained on the MIMIC-III dataset (Johnson et al., 2016).

Figure 1: BiLSTM-CRF neural network architecture



3 Results

On Task 1, our system performed with an F1 score of 0.4317, Precision of 0.3223, and Recall of 0.6534.

On Task 2, on the relaxed metric, our system performed with an F1 score of 0.535, Precision of 0.415, and Recall of 0.753; on the strict metric, our system performed with an F1 score of 0.269, Precision of 0.206, and Recall of 0.390.

4 Conclusion

Overall, our systems for Tasks 1 and 2 consisted of a combination of (1) lexicon selection and domain-specific feature engineering; (2) classical machine learning techniques such as logistic regression; and (3) neural architectures, including BioBERT and BiLSTM-CRF models. We found simpler models consisting of lexicon selection and classical machine learning models (such as the logistic regression model discussed previously) performed better with limited datasets and offered explainability into feature importance. In the Named Entity Recognition task, we utilized a deep learning approach, given the demonstrated effectiveness of such an architecture in this domain (Lample et al., 2016). We expect to improve the performance of our systems through further refinement of our feature engineering and tuning of our model parameters.

Acknowledgments

The research team would like to thank Dr. Chitta Baral of Arizona State University for providing guidance and mentorship for this project.

References

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Samarth Rawal, Ashok Prakash, Soumya Adhya, Sidharth Kulkarni, Saadat Anwar, Chitta Baral, and Murthy V. Devarakonda. 2019. Developing and using special-purpose lexicons for cohort selection from clinical notes. *CoRR*, abs/1902.09674.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012. Modeling spread of disease from social interactions.