# Detection of Adverse Drug Reaction mentions in tweets using ELMo

**Sarah Sarabadani**
Klick Health, Toronto, Canada.
ssarabadani@klick.com

## Abstract

This paper describes the models used by our team in SMM4H 2019 shared task (Weissenbacher et al., 2019). We submitted results for subtasks 1 and 2. For task 1 which aims to detect tweets with Adverse Drug Reaction (ADR) mentions we used ELMo embeddings which is a deep contextualized word representation able to capture both syntactic and semantic characteristics. For task 2, which focuses on extraction of ADR mentions, first the same architecture as task 1 was used to identify whether or not a tweet contains ADR. Then, for tweets positively classified as mentioning ADR, the relevant text span was identified by similarity matching with 3 different lexicon sets.

## 1 Introduction and task description

Twitter is an ever-growing store of daily generated data. Given the huge number of tweets talking about drug-related issues, social media mining is applicable to areas such as pharmacovigilance (Lee et al., 2017; Nikfarjam et al., 2015; Ginn et al., 2014; Freifeld et al., 2014; Bian et al., 2012).

Tasks 1 and 2 focuses on detecting tweets with ADR and identifying location of mentions. We are provided with 25,672 tweets (2,374 positive and 23,298 negative) and approximately 5,000 unlabeled tweets as a validation set. For the second task, a subset of 2,367 tweets from the first task was provided (1,212 positive and 1,155 negative). The evaluation data comprises 1,000 tweets (~500 positive, ~500 negative).

## 2 Preprocessing

Stop words and punctuations were removed from tweets and all drug names found in the FDA's Approved Drug Products list[1] were replaced by the word "drug". Word stemming and tokenization were performed using nltk python library.

## 3 Methods

### 3.1 task 1

For this task, we used 4 deep learning models. The architecture of the first 3 models were relatively similar, differing in the embedding layer.

The first model involves character embedding with dimension equal to the total number of unique characters in training set including emojis. The output of this layer is fed to a series of 6 convolutional neural network layers (CNNs) with ReLU activation. Each CNN used 256 filters, with a filter size of 7 for the first two layers and 3 for the rest. Max pooling with size 3 was used for the first two and last CNNs. The CNNs' output was fed into a bidirectional LSTM (Bi-LSTM) with 2*200 units, whose output was flattened to feed into two dense layers. We used two fully connected layers with 1024 units each, ReLU activation, and dropout of 0.5. Finally, we used a dense layer with size two and softmax activation. We used Adam as the optimizer and binary cross-entropy as the loss function. The model was trained with 10 epochs and batch size of 128.

The second architecture was identical to the first, except the first layer was a word embedding using GloVe[2] pre-trained on Twitter data with embedding dimension of 100.

[1] https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

[2] https://nlp.stanford.edu/projects/glove/



*Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 120–122
Florence, Italy, August 2, 2019. 

The third model was a concatenation of word and character embeddings. We combined the Bi-LSTM output of the first and second models and then applied dense layers as before.

After building the above models, we tried to improve the outcomes by adding layers and features. We used a multi-head self-attention with an attention width of 15 and ReLU activation. We also explored the effect of sentiment features. Since the data classes were imbalanced, we tried to make class sizes equal by downsampling and upsampling. In downsampling, samples from the majority class (tweets without ADR mentions) were randomly sampled without replacement. In upsampling we did the opposite, adding samples from the minority class with replacement. None of these strategies substantially altered our baseline results.

In our final model, we used ELMo (Peters et al., 2018) (Embeddings from Language Models) with 1024 dimensions. In contrast to traditional word embeddings such as GloVe and word2vec, ELMo assigns each word to a vector as a function of the entire sentence containing that word. Therefore, the same word can have different embeddings depending on its context. Since ELMo already captures character-level information under the hood, we decided to encircle the complexity inside the embedding layer and used only two additional dense layers with 256 and 2 units, using ReLU and softmax activations, respectively.

### 3.2 Methods for task 2

To identify the text spans of ADR mentions, first the model developed for task 1 was used to determine whether each tweet mentions an ADR. Then the similarity between each tweet and 3 different lexicon sets (Nikfarjam et al.[3], MedDRA (Medical Dictionary for Regulatory Activities)[4], and CHV (Consumer Health Vocabulary) [5] ) was measured.

To calculate similarity, each tweet and lexicon was converted to a set of word stems. Since similarity measures such as cosine or Jaccard are highly affected by other non-ADR words, we defined similarity as the percent of word stems of a lexicon that exist in a tweet. For each tweet, only lexicons with a 100% match were kept.

## 4 Results, discussion, and next steps

Among all architectures, the best results came from ELMo embedding (F1 = 0.64). Therefore, we only submitted ELMo results with 5, 10, and 15 epochs. The model performed less well for the validation set (F1 = 0.41), below the average F1 score of 0.50 among all teams, which might result from overfitting. Using more sophisticated architecture after the embedding layer might improve performance.

Since task 2's performance depends strongly on task 1, we also scored lower on this task compared to the team average (0.40 vs. 0.54). Since ADR phrases and tweets do not always lexically match, approaches such as named entity recognition (NER) might perform better.

Other approaches to improve performance:

Task 1:

- Try other embeddings such as BERT
- Experiment with more complex architectures after the ELMo layer
- Add part of speech (POS) tags
- Add topic modeling and tweet cluster features

Task 2:

- Search Twitter for keywords from lexicon sets to augment the training set with new tweets which mention ADRs
- Try NER

## Acknowledgment

I would like to thank Maheedhar Kolla who provided insight and expertise that significantly assisted this work.

I would also like to show my gratitude to Peter Leimbigler for comments that greatly improved the manuscript.

Finally, special thanks go to Alfred Whitehead for supporting me to participate in this challenge.

[3] http://diego.asu.edu/Publications/ADRMine.html

[4] https://www.meddra.org/how-to-use/support-documentation/english

[5] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/

## References


Jiang Bian, Umit Topaloglu, and Fan Yu. (2012, October). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing* (pp. 25-32). ACM.

Clark C. Freifeld, John S. Brownstein, Christopher M. Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug safety*, *37*(5), 343-350.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. (2014, May). Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing* (pp. 1-8).

Kathy Lee, Ashequl Qadir, Sadid A. Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. (2017, April). Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 705-714). International World Wide Web Conferences Steering Committee.

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*, *22*(3), 671-681.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, Graciela Gonzalez-Hernandez. 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*