A Pharmacovigilance Application of Social Media Mining: An Ensemble Approach for Automated Classification and Extraction of Drug Names in Tweets.

Luis Alberto Robles Hernandez
Department of Computer Science
Georgia State Univeristy
Atlanta, Georgia
lrobleshernandez1@student.gsu.edu

Rajath Chikkatur Srinivasa
Department of Computer Science
Georgia State University
Atlanta, Georgia
rcs1@student.gsu.edu

Juan M. Banda
Department of Computer Science
Georgia State University
Atlanta, Georgia
jbanda@gsu.edu

Abstract

Researchers have extensively used social media platforms like Twitter for knowledge discovery purposes, as tweets are considered a wealth of information that provides unique insights. Recent developments have further enabled social media mining for various biomedical tasks such as pharmacovigilance. A first step towards identifying a use-case of Twitter for the pharmacovigilance domain is to extract medication/drug terminologies mentioned in the tweets, which is a challenging task due to several reasons. For example, drug mentions in tweets may be incorrectly written, making the identification of these mentions more difficult. In this work, we propose a two step approach, first, we focused on classifying tweets with drug mentions via an ensemble model (containing transformer models), second, we extract drug mentions (along with their span positions) using a text-tagging/dictionary based approach, and a Named Entity Recognition (NER) approach. By comparing these two entity identification approaches, we demonstrate that using only a dictionary-based approach is not enough.

1 Introduction

Twitter, one of the most popular social media platforms in recent years with around 397 million active users[1] has been utilized as "an important source of patient-generated data that can provide unique insights into population health" (Weissenbacher et al. [2]).

Classifying tweets mentioning drug terminologies as well as extracting them is an important research topic since it has many important applications in some areas such as pharmacovigilance [3, 4, 5], as

Preprint. Under review.

well as applications ranging from emotional contagion[6] to toxicovigilance[7]. Furthermore, using a social media like Twitter to perform this task is challenging since the content of the tweets may have misspellings (e.g., “aspirn” for “aspirin”) as well as user-created abbreviations (e.g., “COC” for “Cocaine”). Additionally, some tweets may contain slang terms for drug names (e.g., “Brownies” for “Cannabis”). Moreover, this kind of task related to identifying drug mentions in tweets that are typed by humans with a lot of ambiguities becomes difficult for machine learning models to obtain good results. Davy et al.[2] performed an ensemble approach by training deep learning algorithms to build *Kusuri*, an ensemble learning classifier framework. They were able to achieve a performance close to human annotators with an F1 score of 93.7%

The work done on this paper is part of the shared task from BioCreative[8], which consists of classifying tweets containing drug mentions and extracting medication names from them. To address this task, we proposed an ensemble model to classify these tweets, as well as two approaches (a dictionary-based approach and a Named Entity Recognition BERT-based model) for the extraction of drug mentions from tweets, as well as comparing the performance of these two approaches. This paper is organized as follows: the methodologies implemented in this project, the dataset used for training, the results obtained for the classification and extraction process, and a conclusion of this work.

2 Dataset selection

For the training process, we used the dataset provided by the *BioCreative VII challenge Track 3 - automatic extraction of medication names in tweets*[8], containing around 98,000 tweets (5,200 tweets mentioning at least one drug) for the training set, and 38,000 tweets (with only 105 tweets mentioning medication names in this set) for the validation set. It is important to emphasize that the splits for training and validation were provided by the BioCreative shared task, and the class imbalance (as well as the low count of positive class) was the challenge behind this shared task.

Split group	Tweets with drug mentions	Tweets without drug mentions	Total (tweets)
Training Data	5,209	93,417	98,636
Validation Data	105	38,044	38,149
Total	5,314	131,461	136,775

Table 1. Dataset used for the fine-tuning process

And for the extraction process, the dataset used was the same provided (from Table 1) by Critical Assessment of Information Extraction systems in Biology[8], but only the tweets in the "tweets with drug mentions” class from the training dataset were used (around 5,209 tweets from Table 1), since the NER BERT-Base model system needs to be trained to identify possible drug mentions. Furthermore, a second dataset created by Tekumalla et. al. [9] was included for this process, consisting of a random subset with 190,000 tweets. In addition, the reasons for not using the entire dataset were due to the size of it (containing billions of tweets) and the processing time to feed the NER system. Hence, 195,209 tweets (from both datasets) were used for this process (as shown in Table 2).

Group	Tweets with drug mentions
BioCreative’s dataset[8]	5,209
Twitter dataset with drug mentions[9]	190,000
Total	195,209

Table 2. Dataset used for the drug name extraction process

From the previous data, a pre-processing was implemented by removing URLs, mentions (i.e. "@SomeUser"), and emojis by using the Social Media Mining Toolkit (SMMT) by Tekumalla et al.[10].

3 Methodology

For project implementation, we undertook a multi-step approach where we divided the project into the following five main categories to make development purposes smoother:

3.1 Drug slang dictionary creation

For the first step, a dictionary with a mapping of drug names to slang terms was needed. The primary goal with drug slang terms data extraction was to have a unique list of most probable used drug slang. Dictionary creation process was started by extracting drug slang names from different sources. We identified five sources[11, 12, 13, 14, 15], both official and non-official, which maintained a mapping of list of drugs to their widely used slang in various formats (HTML, and PDF). Python based Tabula and BeautifulSoup libraries were used to web scrape these sources.

Around 2,500 drug slang were extracted from the previous sources, although some slang terms can have multiple meaning, and this could lead to an ambiguity, thus resulting in improper extraction of spans for the last step (extraction of drug mentions from tweets). In general, some cases of ambiguous terms may include Words in other languages, terms related to other domains (Numbers, acronyms, etc.), words with two characters length or less, and/or names.

Manual disambiguation was performed in the dictionary by deleting terms that can have multiple meanings such as: a name, a number, a word with 2 characters length or less, a word in a different language, and/or an acronym. During this process, we encountered some ambiguous terms (i.e. *"ercs"*,*"rims"*, *"paulas"*,*"tires"*, *"buttons"*, etc.) in which they can refer to something else (such as an object, a name, a number, or an acronym). After manually removing all possible ambiguous terms, the drug slang dictionary was reduced to only around 900 drug slang.

3.2 Fine-tuning and creation of ensemble model

For the fine-tuning process, we used a dataset (as shown in Table 1 from Dataset selection) provided by Critical Assessment of Information Extraction systems in Biology[8]. The following pre-trained models were fine-tuned: BERT Uncased[16], BioBERT[17], and CT-BERT[18]. For the previous transformer models, the following parameters were used:

- **Dataset split:** 72.11% for training and 27.89% for validation (As seen in Table 1).
- **Number of epochs:** 3 epochs.
- **Max length:** 300 (The limit of characters for a tweet is 280. However, the max length was set to 300 due to some special tokens like [SEP] or [CLS] added for each tweet, thus making it longer).
- **Learning rate:** 2e-5

The number of epochs and learning rate specified for this fine-tuning process are recommended by the authors from the previously listed models. For example, Devlin et. al.[16] performed several tasks on different models (such as BERT-base and BERT-Large) by specifying different number of epochs and learning rate. They concluded that specific hyperparameters worked well across all tasks, which are the following ones:

- **Learning rate (Adam):** 5e-5, 3e-5, or 2e-5.
- **Number of epochs:** 2,3, or 4.

Also, with any of the previous hyperparameters, according to Devlin et. al.[16], large datasets (with more than 100K examples) are less sensitive than small datasets.

With all the models trained, an **ensemble approach** was implemented for the **validation process** to improve every single fine-tuned transformer model (BERT, CT-BERT, and BioBERT). The way in which this process was carried out was considering the following: a weight was given for each of the 3 fine-tuned models by using the F1-score obtained from the "Tweets with drug mentions" class. If a specific model predicted a tweet containing drug names, a value of 1 was assigned, otherwise a value of -1 was given.

Based on the weight (f1-score) and predictions (either 1 or -1) for each model, the following **formula** was used to get the final predictions for the ensemble model:

$$final\ prediction = \sum_{n=1}^{3}(f1\ score\ model\ n) * (prediction\ value\ model\ n)$$

Based on the numbers, if the **final prediction** was greater or equal than zero, it was classified by the ensemble model as a Tweet with drug mentions, otherwise, it was classified as a Tweet without drug mentions.

3.3 Extraction

Once classifying all the tweets, the next step we undertook was implementing a process to extract the drug mentions from the tweets that were classified as "tweets with drug mentions" by the ensemble model. For this process, we used to different approaches: the first one consisted only of a dictionary-based approach, while the second one consisted of using a dictionary and a NER BERT-based model. The main reason we created two different approaches, was to demonstrate that using only a dictionary to extract drug mentions from a tweet was not enough, since drug mentions in tweets may contain misspellings. The process we done for each approach is as follows.

3.3.1 Extraction - Using a drug slang dictionary

Using the dictionary previously created with all the drug slang terms, an extraction process was implemented (by matching them using regular expressions) to those tweets that the ensemble model classified as "tweets with drug mentions".

The output generated from the extraction process consisted of a Tab Separated Value (TSV) file containing (for each drug mention found) the ID of the tweet, the body of the tweet, as well as the extracted drug mention along with their span position (start and ending).

3.3.2 Extraction - Using a Named Entity Recognition BERT model

The other approach used for the extraction process consisted in fine-tuning a Named Entity Recognition (NER) BERT-based model in order to extract possible drug mentions in tweets classified by the ensemble model as "tweets with drug mentions".

The dataset for the extraction process (using the NER BERT-based approach) was used to fine-tune the NER BERT-based NER model. Moreover, with the drug slang dictionary created before, drug terms from the dictionary created by Tekumalla et al.[9] (containing around 20,000 terms) were also added to our dictionary of slang terms (which originally had around 900 drug slang). Furthermore, in order to fine-tune the previous model, a tokenization and tagging process was needed. The process consisted as follows:

- Using the filtered dataset, we tagged each token as "DRUG" from the tweet if it matched with one of the drug terms from the merged dictionary. (As seen in Figure 1).
- If a token didn't match with any of the terms from the dictionary, they were tagged with an "O" (meaning as "Others") as seen in Figure 1 and Figure 2.
- If a drug mention contained multiple tokens, a "I-DRUG" token was placed in the first token, and a "I-DRUG" token was placed for the rest of the tokens for that drug mention (As seen in Figure 2).
- After the dataset was tokenized and tagged, it was divided as follows: 80% of the tagged tokens were used for training and 20% for validation (randomly selected).

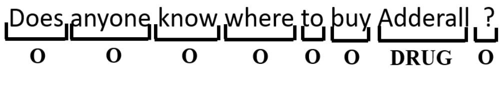

Figure 1: Example of a tagged tokenized tweet (including a one-word drug name)

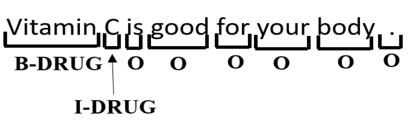

Figure 2: Example of a tagged tokenized tweet (including a multi-word drug name)

For the fine-tuning process, the following parameters were used: 80% for training and 20% for validation, 3 epochs, max length of 300 characters, (The limit of characters for a tweet is 280. However, the max length was set to 300 due to some special tokens like [SEP] or [CLS] added for each tweet, thus making it longer), and a learning rate of 2e-5. The number of epochs and learning rate specified for this process are recommended by the author of this model (as explaned in section 3.2). Before the fine-tuning process, a pre-processing to the dataset was implemented by removing URLs, mentions (i.e. "@SomeUser"), and emojis by using the Social Media Mining Toolkit (SMMT) by Tekumalla et al.[10].

Once fine-tuned the BERT-base model, we used it to identify the drug mentions, by tagging each token from a given tweet (as seen in Figure 1 and Figure 2). Once each token was tagged, we extracted their span positions as well as the drug mention(s) in a TSV (Tab Separated Value) file.

4 Results

4.1 Ensemble model performance

As mentioned before, using an ensemble approach with the previous formula explained, would lead us to obtain favorable results for the minority class (tweets with drug mentions), specifically for the F1-score. Moreover, by looking at the performance metrics obtained from the ensemble model in Figure 3, we can see that it achieved an F1-Score of 0.8818, surpassing the results obtained from every single fine-tuned transformer model.

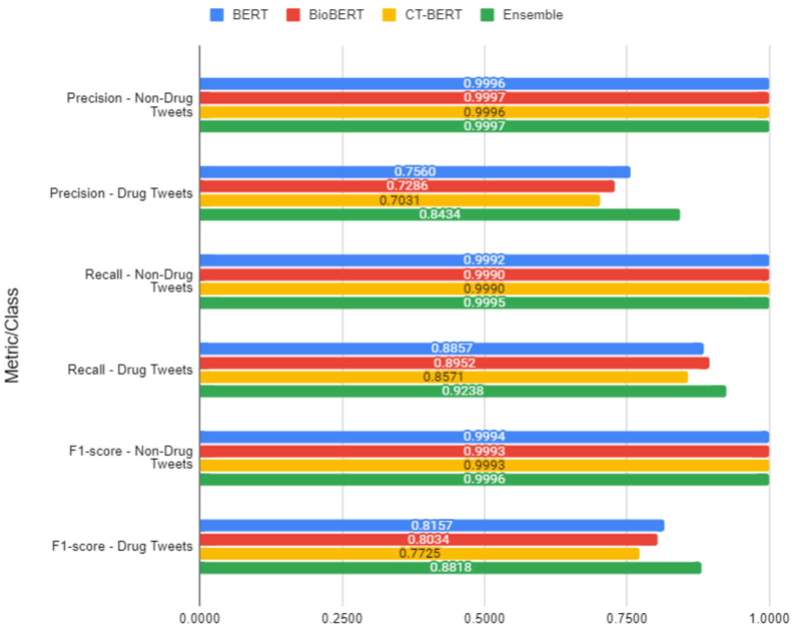

Figure 3: Single model vs Ensemble model performance comparison

4.2 Prediction results

Out of the 105 tweets containing drug mentions from the validation dataset, if we look at the confusion matrix from Table 2, the ensemble model was able to correctly classify 97 tweets in the "tweets with drug mentions" class (or true positives), and only 8 tweets that actually contained a drug mentions, were incorrectly classified in the "tweets without drug mentions" class (or false negatives). Also, 18 tweets that actually doesn't contain drug mentions, were classified in the "tweets with drug mentions" class (false positives).

	Tweets without drug mentions	Tweets with drug mentions
Tweets without drug mentions	38,026	18
Tweets with drug mentions	8	97

Table 2. Confusion matrix from ensemble model

4.3 Extraction results

After the extraction process on the validation dataset from Table 2, using the first approach (dictionary with drug slang terms) we were able to extract 9 drug mentions from 9 tweets. On the other side, using the second approach (using the NER BERT-based model and the dictionary with drug slang terms and drug terms from the dictionary created by Tekumalla et al.[9]) we were able to extract 85 drug mentions out of 85 tweets. The gold standard (validation dataset) contains 105 drug mentions on 105 tweets (one drug mention per tweet). As shown in Figure 4, we can observe the performance obtained for each approach including the drug terms that were not extracted.

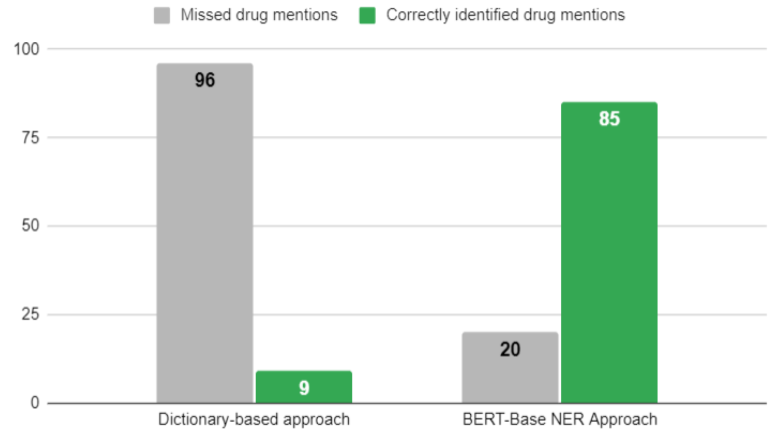

Figure 4: Extraction performance comparison - Dictionary-based approach vs BERT-based approach

Additionally, we can see that the second approach (using a Named Entity Recognition BERT-based model) performed a lot better than the second approach when extracting drug mentions in the validation dataset. Therefore, we ended up using this approach for the extraction process after looking at the results obtained.

5 Conclusions

By creating automated approaches we can remove labor intensive manual curation, thus making the process faster, especially from sources like Twitter which contains a huge amount of data to handle and process. Furthermore, despite the very imbalanced classes presented in the training and validation dataset when implementing this automated approach, the ensemble model was able to perform better than any single fine-tuned model, specifically for the F1-score obtained.

The process of drug extraction was not simple or straightforward as one needs to consider various factors like occurrences of more than one drug mention in a single tweets, or misspelled drug mentions, for example. The disambiguation process was challenging as it involved words from different language and words with multiple meanings, thus making the manual disambiguation difficult. The second approach implemented by using a Named Entity Recognition BERT model, performed a lot better than using only a dictionary with drug slang terms, specifically because the dictionary used not only included the drug slang terms, but also drug terms from the dictionary created by Tekumalla et al.[9].

To improve the results obtained from the extraction process, additional steps can be done in order to extract even more drug mentions, like for example, including in the dictionary possible misspellings a drug term may have (since Twitter is a social media in which we can find possible typos in the content of a specific tweet). Therefore, a Keyboard distance approach would be an option that can be implemented in order to improve the results obtained when extracting drug mentions from a tweet.

References

[1] Most used social media 2021. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed: 2021-9-2.

[2] Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez. Deep neural networks ensemble for detecting medication mentions in tweets. *J. Am. Med. Inform. Assoc.*, 26(12):1618–1626, December 2019.

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

[4] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, June 2012.

[5] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.*, 53:196–207, February 2015.

[6] Cristina Crocamo, Marco Viviani, Lorenzo Famiglini, Francesco Bartoli, Gabriella Pasi, and Giuseppe Carrà. Surveilling COVID-19 emotional contagion on twitter by sentiment analysis. *Eur. Psychiatry*, 64(1):e17, February 2021.

[7] Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter. *Drug Saf.*, 39(3):231–240, March 2016.

[8] Cecilia Arighi, Martin Krallinger, and Florian Leitner. BioCreative VII track 3 - automatic extraction of medication names in tweets. https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/. Accessed: 2021-3-2.

[9] Ramya Tekumalla, Javad Rafiei Asl, and Juan M Banda. Mining archive.org's twitter stream grab for pharmacovigilance research gold. *ICWSM*, 14:909–917, May 2020.

[10] Ramya Tekumalla and Juan M Banda. Social media mining toolkit (SMMT). *Genomics Inform.*, 18(2):e16, June 2020.

[11] DEA Houston Division. Slang terms and code words: A reference for law enforcement personnel. https://www.dea.gov/sites/default/files/2018-07/DIR-022-18.pdf. Accessed: 2021-3-5.

[12] Drug slang code words. https://www.psychiatryadvisor.com/home/dea-drug-slang-code-words/. Accessed: 2021-3-5.

[13] Glossary of slang drug names. https://www.banyantreatmentcenter.com/facilities/chicago/about/slang-drug-terms-glossary, September 2018. Accessed: 2021-3-5.

[14] Street or slang names for drugs. https://www.snohd.org/DocumentCenter/View/2516/Drug_Names_Slang_2019_05_09?bidId=, May 2019. Accessed: 2021-3-5.

[15] T Buddy. Common slang terms for different types of drugs. https://www.verywellmind.com/glossary-of-drug-related-slang-terms-67907. Accessed: 2021-3-5.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. January 2019.

[18] Martin Müller, Marcel Salathé, and Per E Kummervold. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. May 2020.