An ensemble approach for classification and extraction of drug mentions in Tweets

Luis Alberto Robles Hernandez, Rajath Chikkatur Srinivasa, Juan M. Banda
Department of Computer Science, Georgia State University, Atlanta, Georgia, USA

***Abstract*— Twitter, one of the most popular social media sites in recent years, has been considered a unique source to provide insights in areas like pharmacovigilance or biomedical studies. One of the main issues is that social media is informal, meaning that the content provided by users may contain some misspellings, and identifying relevant entities is very challenging to perform in some cases. To address this problem, in the context of identification of medication mentions, we trained an ensemble model to classify tweets that may contain drug mentions, and a fine-tuned Named Entity Recognition BERT-based model to extract identified mentions in relevant tweets. An additional challenge with the dataset provided is the high imbalance between classes. Despite these drawbacks, we were able to extract a high number of drug mentions from the validation dataset of tweets, and demonstrated that using an ensemble model to classify tweets performed better than using any single model used for this work.**

Keywords—Social media mining, pharmacovigilance, information retrieval, knowledge discovery, named entity recognition

I. INTRODUCTION

With around 397 million active users as of July 2021, Twitter is one of the most used social platforms around the world (1). Also, this social media has been utilized as an important source of patient-generated data that “can provide unique insights into population health” (2). The extraction of drug mentions from tweets is an important research topic, especially for the pharmacovigilance area (3–5), which is related to the detection, as well as the prevention of adverse effects of drugs.

Two of the main issues to address in detecting drug mentions from this social media is the effectiveness of extracting drug mentions even if they contain misspellings, or if a tweet contains common slang names for drugs (i.e. “Moon gas” for “Inhalants”). To address this problem, in this project we proposed an ensemble approach to classify tweets that may contain drug mentions, as well as implementing a Named Entity Recognition (NER) model to detect and extract the span positions of drug names from tweets classified (using an ensemble model) as tweets containing drug mentions. This project is part of the BioCreative Track VII Task 3 competition (6). Also, this paper is organized as follows: the methodologies used for this project, the results obtained, a comparison between the average from all the participants obtained from this task, and a conclusion of this work.

II. METHODOLOGY

During the implementation process, several steps were completed and divided into the following main categories:

A. *Fine-tuning process for classification of tweets*

Before performing the drug name extraction from tweets, a model to classify tweets with or without drug mentions was needed. Therefore, an ensemble approach was performed by fine-tuning the following transformer models:

- BERT (7)
- CT-BERT (COVID Twitter BERT) (8)
- BioBERT (9)

The dataset used for the fine-tuning process was provided by Critical Assessment of Information Extraction Systems in Biology (6), which contains the following data:

TABLE I. DATASET USED TO CREATE THE ENSEMBLE MODEL

Group	*Drug tweets*	*Non-drug tweets*	*Total (tweets)*
Training	5,209	93,417	98,626
Validation	105	38,044	38,149
Total	5,314	131,461	**136,775**

As indicated in Table I, the classes are imbalanced, which may lead to a low accuracy, precision and recall (specifically for the tweets with drug mentions). Therefore, by using an ensemble model, these scores may improve compared to only using a single model, which will be explained later in this document. As a first step, the dataset from the previous table was categorized by assigning a value of 1 for the positive class (drug tweets), and a value of 0 for the negative class (non-drug tweets).

The previously listed transformer models (BERT, CT-BERT, and BioBERT) were fine-tuned by using the following parameters:

- **Dataset split (from Table I):** 72.11% for training and 27.89% for validation (to obtain the performance metrics such as accuracy, recall, precision, and f1-score).
- **Number of epochs:** 3 epochs.

- **Learning rate:** 2e-5.
- **Max length:** 300 characters (The limit of characters from a tweet is 280. However, when adding special tokens such as [SEP] and/or [CLS] the length can be longer)

Before the fine-tuning process, a preprocessing step was performed on each tweet by removing URLs and mentions (i.e. “@User”), as well as emojis by using the Social Media Mining Toolkit (10).

To build the ensemble model, the F1-scores from the “tweets with drug names” class were considered since we are focusing on detecting which tweets can include drug mentions, hence the following formula was used to get the final prediction for the ensemble model:

$$final\ prediction = \sum_{k=1}^{3} f1\_score\_model\_k * final\_prediction\_model\_k$$

Given the previous formula, *f1_score_model_k*, represents the f1-score obtained from each of the three fine-tuned transformer models, while *final_prediction_model_k* represents the prediction obtained from a specific model (-1 if a tweet was classified as a tweet without drug mentions, and 1 if a tweet was classified as a tweet with drug mention(s)). Based on the previous formula, if the *final prediction* for a specific tweet was greater or equal than zero, it was classified as a tweet with drug mentions, otherwise it was classified as a tweet without drug mentions.

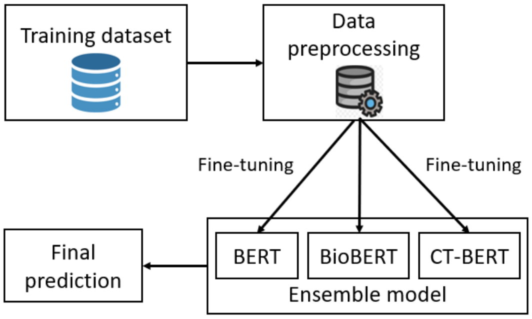

Fig. 1. Fine-tuning process and creation of the ensemble model

B. *Fine-tuning process for extraction of drug mentions in tweets*

After classifying the tweets in the positive or negative class, a pre-built NER model was fine-tuned in order to extract drug mentions from tweets classified in the positive class. As shown in Table II, the dataset was obtained from two corpora: Critical Assessment of Information Extraction Systems in Biology (6), and Social Media Mining for Health Applications (SMM4H´18). It is important to emphasize that only the tweets from the positive (tweets containing drug mentions) class were used to fine-tune the BERT extractor. Additionally, data from the Twitter dataset with drug mentions (11) in which around 190,000 tweets were included as well. This dataset is a subset from the original dataset(11) containing tweets filtered by using a dictionary of drug terms in order to obtain only tweets mentioning medication names. This dictionary was built by manually curating RxNorm and removing ambiguous and very long terms.

Since the datasets used are not annotated, we leveraged the drug and drug slang mention dictionaries to annotate drug mentions in them. This was done in order to be able to fine-tune the NER BERT extractor model.

TABLE II. DATASET USED FOR THE DRUG NAME EXTRACTION

Group	*Tweets with drug mentions*
BioCreative’s dataset (6)	5,209
Twitter dataset with drug mentions (11)	190,000
Total	195,209

The previous model was fine-tuned by using the following parameters:

- **Dataset split:** 80% for training and 20% for validation.
- **Number of epochs:** 3 epochs.
- **Learning rate:** 2e-5.
- **Max length:** 300 characters (The limit of characters from a tweet is 280. However, when adding special tokens such as [SEP] and/or [CLS] the length can be longer)

Before the fine-tuning process, a preprocessing step was performed on each tweet by removing URLs and mentions (i.e. “@User”), as well as emojis by using the Social Media Mining Toolkit (10). Additionally, the dataset for the NER model was required to be annotated for each token. Therefore, the following steps were done:

- A dictionary of common slang names for drugs was created by extracting them from a number of sources(12–16) and using Python-based *Tabula* and *BeautifulSoup* libraries. To merge all these common slang names obtained from the previous sources and to normalize the dictionary, the following steps were undertaken:
  - A unique id was assigned for every unique slang term.
  - All duplicated drug slang terms were removed, keeping the first occurrence only.
  - All the content of the dictionary was lower-cased.
- Since some terms from the dictionary created had multiple meanings, ambiguous terms were manually removed by considering the following criteria:
  - Words in other languages
  - Terms related to other domains (Numbers, acronyms, etc.)
  - Words with two characters length or less

- The resulting dictionary consisted of slang terms (with around 900 terms) which were not considered as ambiguous (considering the previous criteria). Furthermore, an extra dictionary of drug names(11) was also merged with the previous dictionary obtaining around 20,000 common slang names for drugs and drug names, which was used to prepare the training dataset for the fine-tuning process for the NER BERT-based model.

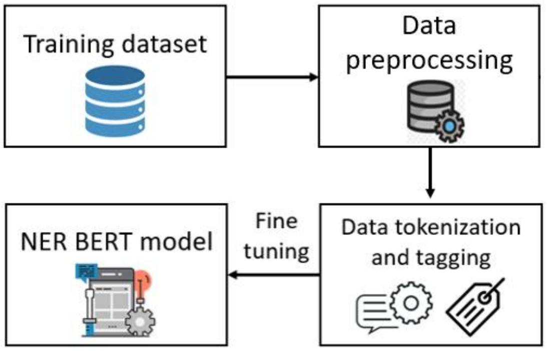

Fig. 2. Fine-tuning process for the NER BERT-based model

- A special delimiter was added for each drug mention found in a tweet (before and after the drug name).

- All the previous tweets were tokenized and grouped per sentence.

- All the previous tokenized tweets were tagged by using the following criteria:

  - If a token was not inside the special delimiter, it was tagged as "O" (as seen in Figure 3 and Figure 4).

  - If a token was inside the special delimiter:

    - If there were multiple tokens (a multi-word drug name), a "B-DRUG" was tagged for the first token, and a "I-DRUG" for the rest of the tokens inside the special delimiter (as seen in Figure 3).

    - If only one token was found inside the special delimiter, it was tagged as "DRUG" (as seen in Figure 4).

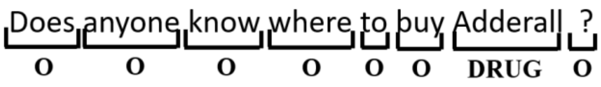

Fig. 3. Example of a tagged tokenized tweet (including a one-word drug mention)

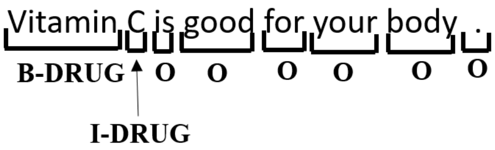

Fig. 4. Example of a tagged tokenized tweet (including a multi-word drug mention)

Once the NER BERT-based model was fine-tuned, we used the results obtained from the ensemble model to extract the drug names for only those tweets that were classified as "tweets with drug mentions".

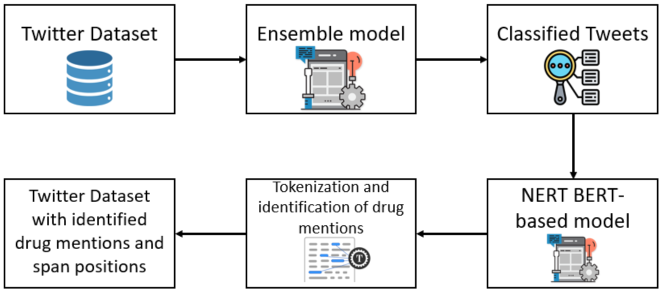
Twitter Dataset
Ensemble model
Classified Tweets
NERT BERT-based model
Tokenization and identification of drug mentions
Twitter Dataset with identified drug mentions and span positions

Fig. 5. Overall process of classification and extraction of drug mentions from tweets

III. RESULTS

For the classification process, in the following plots we compare the performance obtained in the validation dataset (in terms of precision, recall, and f1-score) individually for each fine-tuned transformer model as well as the ensemble model.

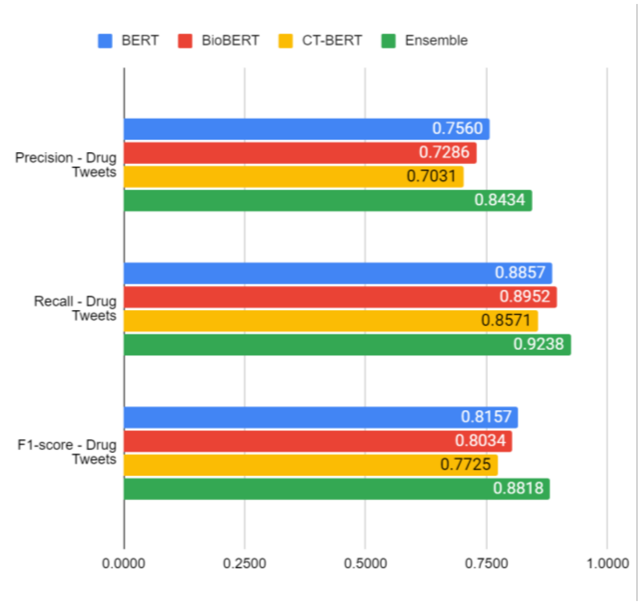

Fig. 6. Single model vs ensemble model performance comparison

As seen in Figure 6, we can observe that the ensemble model obtained the best F1-score when comparing to every individual model. It is important to mention that this ensemble model outperformed or achieved similar scores compared to some classifiers competing in previous Social Media Mining For Health (SMM4H) shared tasks. For example, the ensemble BERT fine-tuned by Dang et. al. (17), achieved a 0.8955 F1-score when classifying tweets in the positive or negative class, compared to the 0.8818 obtained in our ensemble model. Also, the model proposed by Prakash et. al. (18) achieved an F1-score of 0.7356, surpassing the score obtained from our model. Same case with the models trained by Mehnaz et. al. (19) in which the best model (BioMed-RoBERTa) achieved an F1-score of 0.8500.

For the extraction process in the validation dataset, by comparing the results obtained with the gold standard dataset, the final score obtained using the Codalab platform (20) was around 0.72 (where 1 means that all drug mentions were correctly extracted).

Furthermore, the prediction and extraction process were also performed in a test dataset provided by Critical Assessment of Information Extraction Systems in Biology (6) containing 54,482 tweets. The results obtained from this dataset were as follows:

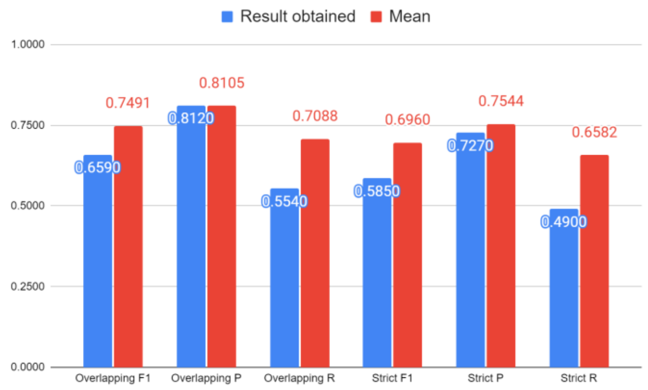

Fig. 7. Comparison between the results individually obtained vs the average obtained from all participants for this task.

As seen in Figure 7, we can highlight that despite the imbalanced dataset for the classification process, we obtained an overlapping precision above 81%, and a strict precision above 72%. Moreover, comparing the metrics obtained to the mean (from all the participants of this task), almost all of them were below the average, with the exception of the overlapping precision which achieved a higher score than the average.

IV. CONCLUSIONS

Looking at the performance obtained from the validation dataset, the classification of tweets using an ensemble model achieved a decent performance in the validation dataset with an F1-score above 0.88. Additionally, when comparing the individual results from each transformer model against the ensemble model, we observed that the second one performed better than any single fine-tuned model. Furthermore, the disambiguation of drug slang for the dictionary was challenging since some terms could have various meanings. Without a disambiguation process, the NER BERT-based model could perform worse.

Moreover, the results obtained from the test dataset were slightly below the average for all the participants in this task, meaning that additional tasks can be performed to improve this result. One of the possible additional tasks that can be done to improve the results is by increasing the number of drug terms from the dictionary (ideally from several official sources), so the NER BERT-based model can recognize and identify even more drug mentions from a tweet.

REFERENCES

1. Most used social media 2021, (n.d.). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed September 2, 2021).

2. D. Weissenbacher, A. Sarker, A. Klein, K. O'Connor, A. Magge, G. Gonzalez-Hernandez, Deep neural networks ensemble for detecting medication mentions in tweets, J. Am. Med. Inform. Assoc. 26 (2019) 1618–1626.

3. D. Weissenbacher, A. Sarker, M.J. Paul, G. 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, Association for Computational Linguistics, Brussels, Belgium, 2018: pp. 13–16.

4. T. Rocktäschel, M. Weidlich, U. Leser, ChemSpot: a hybrid system for chemical named entity recognition, Bioinformatics. 28 (2012) 1633–1640.

5. A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J. Biomed. Inform. 53 (2015) 196–207.

6. C. Arighi, M. Krallinger, F. Leitner, BioCreative VII Track 3 - Automatic extraction of medication names in tweets, (n.d.). https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/ (accessed March 2, 2021)

7. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv [cs.CL]. (2018). http://arxiv.org/abs/1810.04805.

8. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, arXiv [cs.CL]. (2019). http://arxiv.org/abs/1901.08746.

9. M. Müller, M. Salathé, P.E. Kummervold, COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter, arXiv [cs.CL]. (2020). http://arxiv.org/abs/2005.07503.

10. R. Tekumalla, J.M. Banda, Social Media Mining Toolkit (SMMT), Genomics Inform. 18 (2020) e16.

11. R. Tekumalla, J.R. Asl, J.M. Banda, Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold, ICWSM. 14 (2020) 909–917.

12. D.D.I. Intelligence, Slang terms and code words: A reference for law enforcement personnel, (n.d.). https://www.dea.gov/sites/default/files/2018-07/DIR-022-18.pdf (accessed March 5, 2021).

13. Drug Slang Code Words, (n.d.). https://www.psychiatryadvisor.com/home/dea-drug-slang-code-words/ (accessed March 5, 2021).

14. Glossary of Slang Drug Names, (2018). https://www.banyantreatmentcenter.com/facilities/chicago/about/slang-drug-terms-glossary (accessed March 5, 2021).

15. Street or Slang Names for Drugs, (2019). https://www.snohd.org/DocumentCenter/View/2516/Drug_Names_Slang_2019_05_09?bidId=.

16. T. Buddy, Common Slang Terms for Different Types of Drugs, (n.d.). https://www.verywellmind.com/glossary-of-drug-related-slang-terms-67907 (accessed March 5, 2021).

17. H. Dang, K. Lee, S. Henry, Ö. Uzuner, Ensemble BERT for Classifying Medication-mentioning Tweets, in: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Association for Computational Linguistics, Barcelona, Spain (Online), 2020: pp. 37–41.

18. Y. Prakash Babu, R. Eswari, Identification of Medication Tweets Using Domain-specific Pre-trained Language Models, in: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Association for Computational Linguistics, Barcelona, Spain (Online), 2020: pp. 128–130.

19. L. Mehnaz, Automatic Classification of Tweets Mentioning a Medication Using Pre-trained Sentence Encoders, in: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Association for Computational Linguistics, Barcelona, Spain (Online), 2020: pp. 150–152.

20. BioCreative'21, Task 3 - Automatic extraction of medication names in tweets, (n.d.). https://competitions.codalab.org/competitions/23925 (accessed August 30, 2021).