An Enhanced Approach to Identify and Extract Medication Mentions in Tweets via Weak Supervision

Ramya Tekumalla, Juan M Banda
Department of Computer Science, Georgia State University, Atlanta, Georgia, USA

***Abstract*—This paper presents our system details and results of our participation in the Biocreative 2021 Track 3. This track aims at extracting medication mentions from tweets and provides an opportunity for participants to utilize methods beyond lexical matching. In this task, we utilize a weak supervision approach and train several machine learning models with additional data beyond the provided training data. We tested our models using the validation data and submitted two best results. Our best system achieved 0.687 strict F1 and 0.771 strict Precision scores.**

Keywords— Classification, Information Retrieval, Pharmacovigilance, Social Media Mining, Twitter

I. Introduction

Twitter contains an abundance of drug data as users tend to share their experience on social media (1). In the past, several techniques like lexical matching (2,3), language models (4), and frameworks like Kusuri (5) presented successful results in identifying drug tweets from a corpus of tweets. However, all the techniques utilized a supervised learning approach and used an annotated corpus to train the machine learning models. Weak supervision utilizes noisy or limited sources to provide supervision signals for labeling large amounts of training data in a supervised learning setting (6). In this task, we utilized a weak supervision approach to identify the tweets with medication mentions and used the Social Media Mining Toolkit (SMMT)'s NER utility (7) along with a dictionary of drug terms to extract the medication mentioned in the tweet text.

II. Methodology

A. Data Collection

A crucial challenge in this task was to use a highly imbalanced dataset. The training data provided by the organizers contains 84,815 tweets with only 218 tweets mentioning at least one drug. If a model is trained on highly imbalanced data, it would either over or underfit the data during the testing phase. To overcome this challenge, we collected data from several publicly available resources to balance the training data.

As part of our previous work, we curated Tekumalla et al. (8), a silver standard dataset using a heuristic approach which contains a dictionary of terms compiled from RxNorm. This dataset only classifies tweets and is not an annotated dataset. The drug dictionary utilized for filtering the tweets was built based on the following conditions

i) A term length must be greater than 3 characters and less than 33 characters

ii) The language of the term must be English

After retrieving all the terms, we further removed some noisy terms (Eg: disk, foam) since they would filter irrelevant tweets. We also removed large chemical compounds such as "2,10,15,19,23-pentahydrosqualene" since the language on Twitter is unstructured and a user would rarely type the whole text without an error. The final drug dictionary consists of 19,643 terms. Using this dictionary, we filtered 4,214,737 drug tweets from the Internet Archive (9) and curated the silver standard dataset. Additionally, we collected tweets from two different datasets i.e Klein et al. (10) and Sarker et al. (4) which are released by the health processing lab at UPENN. All the tweets collected from the additional datasets are labelled as drug tweets since they were collected using drug terms or variants of drug terms. Table I presents the details of the acquired additional data. We hydrated the tweet ids released by the health processing lab and used the tweet text for this task. We did not manually annotate any dataset listed in Table 1. We did not utilize the SMM4H'18 dataset for this task. We have utilized all the datasets in several combinations to train the machine learning models.

TABLE I. Data collection details

Dataset	No of drug tweets available
Sarker et al. (4)	106,559
Tekumalla et al. (8)	4,214,737
Klien et al. (10)	7,215

B. Classification

To differentiate between the medication and non-medication tweets, we trained several machine learning models with the acquired data. In the classical models front, we utilized scikit-learn's library (11) to implement Logistic Regression, SVM, Naive Bayes, Random Forest and Decision Tree models. The TF-IDF vectorizer was used to convert raw tweet text to TF-IDF features and return the document-term matrix which is sent to the model.

In the deep learning models front, we experimented with the "bert-large-uncased" model, which is of 24-layer, 1024-hidden, 16-heads, 340M parameters and trained on lower-cased English Wikipedia text and book corpus (12).

Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) (13), is a domain-specific language representation model pre-trained on large-scale biomedical corpora. The BioBERT model architecture used for our experiment is 12-layer, 768 hidden size, 12-heads, 1M parameters and trained on PubMed baseline abstracts. The final architecture used in Transformers is Robustly Optimized BERT Pretraining Approach (14) which has an improved pre-training procedure over BERT. We used the "roberta-large" model which is of 24-layer, 1024-hidden, 16-heads, 355M parameters using the BERT-large architecture. The simple transformers library (15) was utilized for implementation since it seamlessly works with the Hugging face's transformer models (16). Apart from transformer models, we also experimented with CNN and LSTM models. The keras implementation of models by Text Classification Algorithms (17) was utilized. For both the models, we used Adam Optimizer, Relu Activation function and RedMed (18) embedding model.

We filtered the biocreative training data (BT) based on the "drug" column in the training data and acquired 218 drug tweets. To this data, we added 106,559 tweets from the Sarker et al. (4) dataset, 1,000,000 tweets from the Tekumalla et al. (8) dataset and 7,215 tweets from the Klien et al. (10) dataset. All the tweets collected from the additional datasets are labelled as drug tweets since they were collected using drug terms or variants of drug terms. We did not pre-tag or annotate any of the drug tweets. We utilized the 84,597 non drug tweets from the biocreative training data and randomly selected 84,319 non drug tweets to balance the dataset. Additionally we added 7,433 non drug tweets from our previous work (19). Table II contains the details of the datasets as well as the number of samples used and the results obtained for several experiments performed. We split the training samples and used 75% of the data for training the model and 25% for evaluating the model.

C. Exploratory Analysis

In total, we experimented with 5 classical models and 5 deep learning models. To evaluate our models, we used the validation set released by the organizers which contains 38,150 tweets out of which only 93 tweets mention at least one drug. With our initial experiments, we determined that the performance of deep learning models were superior when compared to classical models. Hence, we further experimented with only deep learning models to identify the tweets with drug mentions. In our experiments, we started with balancing the imbalanced dataset and increased the drug tweets all the way to a million tweets. We experimented with several control sets to identify the model that classifies the validation set. To evaluate the models, we used Precision (P), Recall (R), F-measure (F) and Accuracy (A) metrics. Table II presents details on the control sets and number of tweets used along with the evaluation metrics. We only included the best results for the models in Table II and did not include the results of all the experiments since some experiments yielded less performance than others.

Post analysis, we identified that the performance of the BERT model was superior when compared to other models. Though the BioBERT model was trained on biomedical Pubmed abstracts, it did not achieve the same performance as BERT since the language used in a tweet is unstructured. A decline in the performance is observed when the training samples of Tekumalla et al. (8) dataset are increased beyond 200,000 samples. We utilized only 100,000 samples from Tekumalla et al. (8) for our best model. We utilized the BERT models trained on (BT + Klien et al. (10) + Sarker et al. (4)) and ((BT + Tekumalla et al. (8) +Klien et al. (10)+ Sarker et al. (4)) for our final submissions.

TABLE II. BEST RESULTS FROM EXPLORATORY ANALYSIS

Data Used	# drug samples	#non drug samples	P	R	F	A	Best Model
BT + Tekumalla et al. (8)	84,319	84,319	0.2645	0.8667	0.4053	0.9927	RoBERTa
BT + Tekumalla et al. (8)	100,496	84,319	0.2783	0.8667	0.4213	0.9931	BERT
BT + Tekumalla et al. (8)	200,496	84,319	0.2662	0.7429	0.392	0.9933	BioBERT
BT + Tekumalla et al. (8). + Klien et al. (10)	107,711	91,534	0.2975	0.7905	0.4323	0.9940	BERT
BT + Klien et al. (10) + **Sarker et al.** (4)	**113,992**	**92,030**	**0.8043**	**0.7048**	**0.7513**	**0.9987**	**BERT**
BT + Tekumalla et al. (8) + **Klien et al.** (10)+ **Sarker et al.** (4)	**213,992**	**92,030**	**0.3359**	**0.8381**	**0.4796**	**0.9948**	**BERT**

D. Entity Extraction

For our final submissions, we retrained the BERT models adding the biocreative validation dataset to the training datasets (BT + Klien et al. (10) + Sarker et al. (4)) and (BT + Tekumalla et al. (8) +Klien et al. (10)+ Sarker et al. (4)). Upon the release of the test set, we used the trained models and acquired predictions for the test set. The primary task of this track is to extract the drug mentions from the tweets and obtain the spans of the drug mention. To complete the task, we utilized the SMMT NER utility which utilizes Spacy (20) library to tag drug mentions in the tweet text from a given dictionary. We utilized the drug dictionary with the SMMT NER utility to tag the drug terms. Since the training data contains a number of terms which are not available in Rxnorm (Eg: birth control), we computed a list of drug terms from the training and validation data and added it to our dictionary. We extracted the drug terms for all the tweets that were classified as a drug tweet by the machine learning model.

III. Results

A. Results from Biocreative task

Table III presents the results received for our two submissions. Our results are compared with the mean of all the results submitted in the competition. The BERT model trained with (BT + Klien et al. (10) + Sarker et al. (4)) datasets achieved better results when compared to our second submission which was trained on (BT + Tekumalla et al. (8) + Klien et al. (10) + Sarker et al. (4)) dataset. Our best submission achieved an **overlapping F1 score of 0.737** and **strict F1 0.687**.

TABLE III. Results of the biocreative task

	Overlapping F	**Overlapping P**	**Overlapping R**	**Strict F**	**Strict P**	**Strict R**
BioCreative Task Mean	0.7491	0.8105	0.7088	0.6960	0.7544	0.6582
(BT + Klien et al. (10) **+ Sarker et al.** (4)	**0.7370**	**0.8310**	**0.6620**	**0.6870**	**0.7710**	**0.6190**
BT + Tekumalla et al. (8) +Klien et al. (10)+ Sarker et al. (4)	0.5180	0.4010	0.7300	0.4810	0.3720	0.680

B. Discussion

The drug dictionary utilized for extracting the spans of the drug term was curated from RxNorm. RxNorm does not contain common drug slang or colloquial terms (Eg: birth control, epidural, pills). The following tweets contain terms which are not available in RxNorm. The tweets are paraphrased due to Twitter's terms and conditions.

i) i'm so scared to stop taking my **birth control**!! thank god it was just a false alarm i don't want to get pregnant any time soon

ii) tonight, i dont have my nausea **pills**... so i'm probably going to die tomorrow…

iii) bloody toothache kept me awake all night.. **painkillers** aren't working, swirling brandy isn't working. i need sleep

For over 40 tweets in the test set, we could not extract the drug span since the term was not available in the dictionary. For all such tweets, we marked the tweets as non-drug tweets even though the machine learning model predicted the tweet to be a drug tweet. The 40 tweets could be either false positives or the text could have had a drug slang term which was not available in our drug dictionary.

Our previous work with weak supervision demonstrated successful results on a classification task (identifying drug mentions from Twitter) with noisy data. However, for this task, the results depict a decline in performance metrics when noisy data is increased. Since the Tekumalla et al. (8) dataset is not limited to a certain number of drug terms, we believe that either the task is limited for certain types of signals (e.g. pregnancy drug/ slang terms) or the amount of noisy data induced during the training phase was higher resulting in a decline in the performance.

C. Summary

To summarize our system, we first collected several datasets (Tekumalla et al. (8) + Klien et al. (10)+ Sarker et al. (4)) in addition to the BT dataset and labelled all the

tweets from collected datasets as drug tweets. We used several training sizes by incrementally increasing the samples of drug tweets in the datasets and trained several machine learning models in a binary classification setting. We tested our models based on the biocreative validation data and used the best models trained on (BT + Klien et al. (10) + Sarker et al. (4)) and (BT + Tekumalla et al. (8) + Klien et al. (10) + Sarker et al. (4)) since they obtained the best F-measure. We retrained the models adding the biocreative validation dataset and finally obtained the predictions on the test data. We filtered all the positive predictions and extracted the spans of the drug term using a drug dictionary. Fig. 1 depicts the overview of our system used for this task.

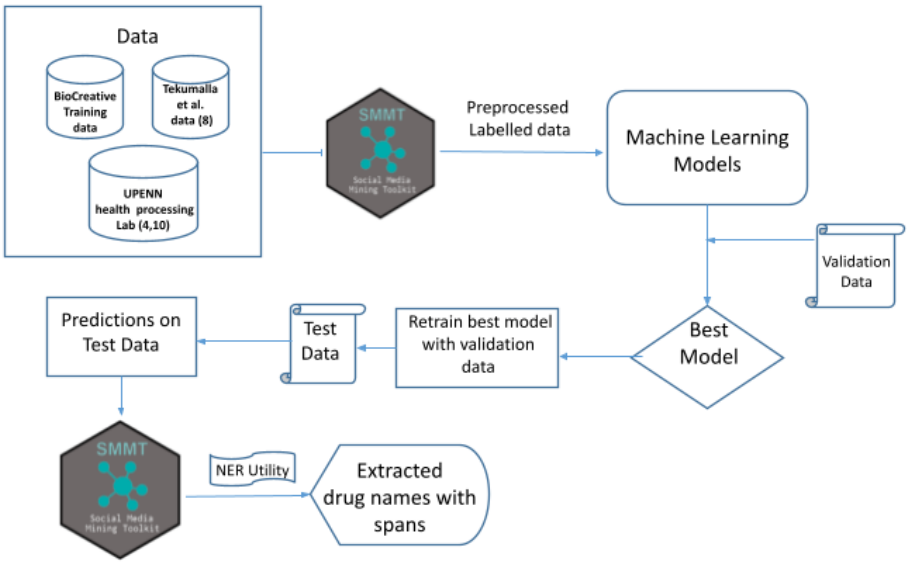

Fig. 1. Overview of the system

IV. CONCLUSION

In this work, we utilized weak supervision to identify tweets with drug mentions and an NER utility to extract the term from the predicted drug tweet. In the future, we would like to use the scispacy for such tasks which contain transformer models trained on biomedical vocabulary and also contain entity linkers to several vocabularies. However, the challenge with using such pipelines is that the common drug slang terms are not available. A possible solution is to identify the common drug slang terms used on Twitter, create a vocabulary, and add it as a pattern in NERs or drug dictionary for entity extraction. Finally, we would like to thank the organizers for this opportunity to demonstrate our methodology on a highly imbalanced dataset.

REFERENCES

1. O'Connor, K., Pimpalkhute, P., Nikfarjam, A., et al. (2014) Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. *AMIA Annu. Symp. Proc.*, **2014**, 924–933.
2. Leaman, R., Wojtulewicz, L., Sullivan, R. C., et al. (2010) Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. .
3. Sarker, A. and Gonzalez, G. (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.*, **53**, 196–207.
4. Sarker, A. and Gonzalez, G. (2017) A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. *Data Brief*, **10**, 122–131.
5. Weissenbacher, D., Sarker, A., Klein, A., et al. (2019) Deep neural networks ensemble for detecting medication mentions in tweets. *J. Am. Med. Inform. Assoc.*, **26**, 1618–1626.
6. Ratner, A., Bach, S., Varma, P., et al. (2019) Weak supervision: the new programming paradigm for machine learning. *Hazy Research. Available via https://dawn. cs*.
7. Tekumalla, R. and Banda, J. M. (2020) Social Media Mining Toolkit (SMMT). *Genomics Inform.*, **18**, e16.
8. Tekumalla, R., Asl, J. R. and Banda, J. M. (2020) Mining Archive. org's Twitter Stream Grab for Pharmacovigilance Research Gold. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14, pp. 909–917.

9. Machine, W. (2015) The Internet Archive. *Searched for http://www. icann. org/icp/icp-1. htm*.
10. Klein, A., Sarker, A., Rouhizadeh, M., et al. (2017) Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system. *BioNLP 2017*, pp. 136–142.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
12. Devlin, J., Chang, M.-W., Lee, K., et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* **(2018)** .
13. Lee, J., Yoon, W., Kim, S., et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
14. Liu, Y., Ott, M., Goyal, N., et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv [cs.CL]* **(2019)** .
15. Rajapakse, T. (2019) simpletransformers. *simpletransformers*; Github, (2019) .
16. Wolf, T., Debut, L., Sanh, V., et al. (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, arXiv:1910.03771.
17. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., et al. (2019) Text Classification Algorithms: A Survey. *Information*, **10**, 150.
18. Lavertu, A. and Altman, R. B. (2019) RedMed: Extending drug lexicons for social media applications. *J. Biomed. Inform.*, **99**, 103307.
19. Tekumalla, R. and Banda, J. M. (2021) Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Computing and Applications*.
20. Explosion, A. I. (2017) spaCy-Industrial-strength Natural Language Processing in Python. *spaCy-Industrial-strength Natural Language Processing in Python*; (2017) .