System Description

Mia Mohammad Imran
Virginia Commonwealth University
Richmond, Virginia, U.S.A.
imranm3@vcu.edu

Abstract

Social media posts have become an enormous source of population health. Users often share their experiences about medication intake and their effects. A fundamental step toward incorporating social media data in pharmacoepidemiologic research is to automatic medication names detection in these posts. However, detecting drug name mentioning posts are challenging as they are often informal and noisy. Medication names are often misspelled and abbreviated. The purpose of this study is to detect medication name spans in tweets.

I. Introduction

Twitter have become a very popular source for information sharing. Automatically detecting tweets which mentions medication names has become an area of interest in recent years. However, tweets are very noisy and informal, and full of misspellings and user-created abbreviations.

In order to facilitate the research on automatic detection of tweets mentioning medication names, BioCreative VII released a task - Automatic extraction of medication names in tweets [1]. This research project is based on the task.

A. Task Definition

The goal of the task is to extract the spans that mention a medication or dietary supplement in tweets. The table I shows an example of the task. If a tweet mentions 2 or more drugs, the tweet is repeated 2 or more times with the mention of each drug in each repetition as shown below. The evaluation data will just contain the tweet IDs and the text of the tweet. The evaluation is based on the normalization task, just the extraction task, i.e. retrieving the span positions. And the evaluation matrics is exact and partial F1-scores for the positive class (i.e., the correct spans of drug name).

Id	Tweet	Has Medication	Begin	End	Span	Drug normalized
397783574797352960	Only 3 Arnica Balms left...	1	8	11	Arnica Balms	arnica balm
404288692514078720	@user sudafed that I'm not sure [..]	1	1	13	sudafed	sudafed
343961712334686205	I like this song!	0	-	-	-	-
424441978835570688	@user no my [..] hydros and moltrin	1	44	49	hydros	hydrocodone
424441978835570688	@user no my [..] hydros and moltrin	1	55	61	moltrin	motrin

TABLE I: Task Description

II. System Description

A. Selecting a Corpus

The dataset in the task was unavailable. So we emailed on of the organizers and they provided an annotated dataset of similar task[2]. This dataset is highly imbalanced. To make it slightly balanced, we collected another dataset used by Alvaro et al. [1]. Both of the dataset were already annotated. We combined the two dataset and generate three datasets for our test purpose. The positive class distribution in these datasets is shown in table II. We also come up with a pre-compiled list of 89 medication and dietary names and their variations based on the provided dataset.

[1]https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-3/

[2]https://competitions.codalab.org/competitions/23705?secret_key=aeda0cb9-a3da-4e4f-9d6a-d639355b455e

Total Entry	Positive-class	Positive-class percentage
52259	200	0.38 %
52559	500	0.95 %
53059	1000	1.88 %

TABLE II: Dataset Distribution

B. Pre-processing

Tweets are often informal. So preprocessing the data is essential. To do preprocessing, we setup some rules:

- Remove names (mentioned with @)
- Remove emojis
- Remove links
- Remove punctuations and non-ASCII characters
- Remove common stop words using Spacy
- Lemmatization
- Filter any tweet that is consist of less than 3 words after above steps.

C. Data

The three datasets are three csv files. The dataset entity structure is: "id, original_tweet, tweet, class". Here, 'id' indicates an unique id; 'original_tweet' indicates the original tweet text; 'tweet' indicates the text after preprocessing; and 'class' indicates whether it contains medication names or not (class value 1 meaning positve class, 0 meaning negative class). For training and test purpose, we split the dataset on 75-25 ratio respectively.

D. Model

In this section, we will describe our system and model. We can divide our system into two steps: 1) Binary classification of a tweet base on whether it contains medication name or not, and, 2) If it contains the medication name, detecting the span. The pipeline of our process can be found in figure 1.

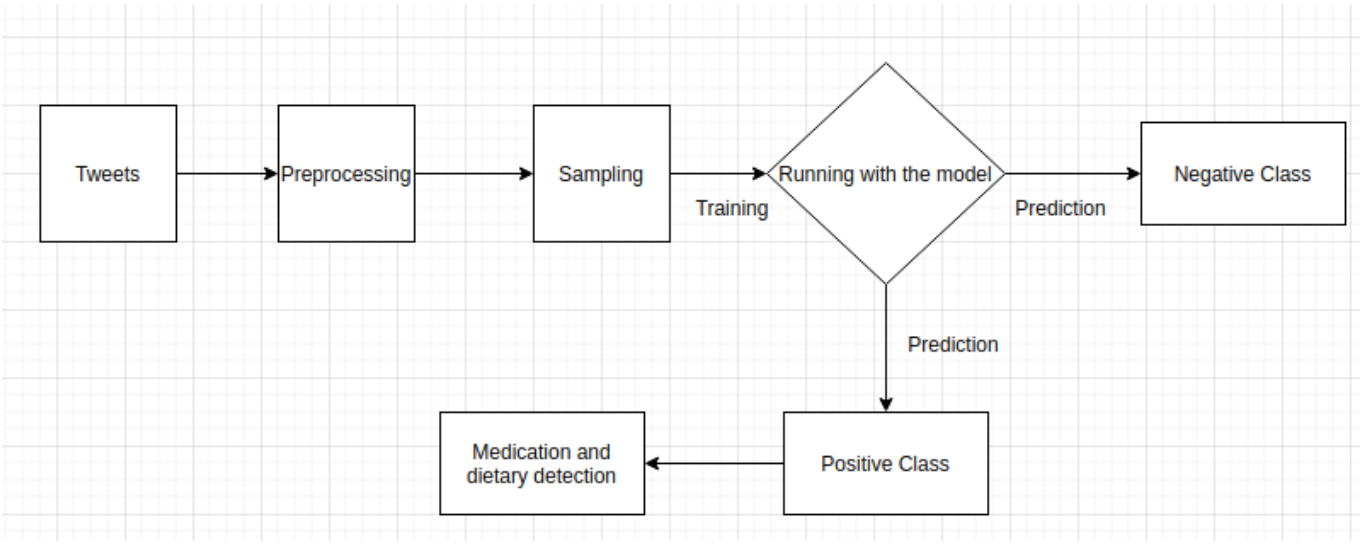

Fig. 1: Pipeline of the system

III. Experiment Setup and Evaluation

In this section, we will present our experiment setup, the evaluation metrics, analyze the results, and do error analysis.

A. Experiment Setup

The model is divided into two parts. We have implemented our model on PyTorch.

1) Tweet Classifier: We have used a Long short-term memory (LSTM) model. LSTM is a variant of Recurrent neural network (RNN). This classifier predicts whether the tweet may contain medication name or not. The model transforms the tweet text into it's vector representation using Glove Twitter word embeddings [3]. We generate vocab of training dataset with minimum frequency three. This model uses two Rectified Linear Unit (ReLU) and one linear-classification for prediction. The model design is shown in figure 2. We have used Adam optimizer and Sigmoid loss function.

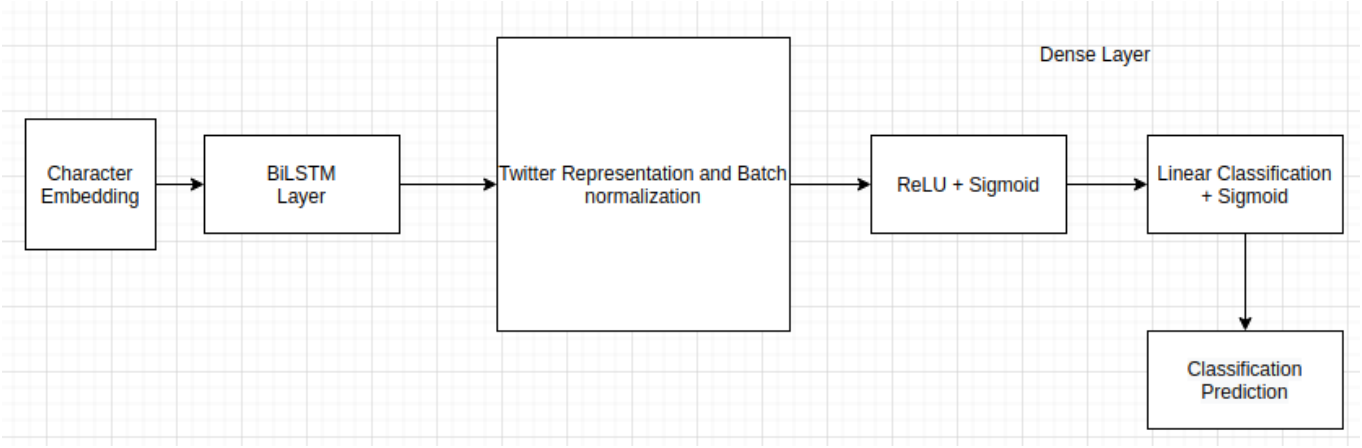

Fig. 2: Architecture of the model

2) Span Detection: For a positive class tweet, we searched a medication name with our pre-compiled medication and dietary list.

B. Evaluation Metrics

BioCreative mentioned f1-score for positive class as a evaluation metric. However, we have added precision, and recall as well for comparison.

C. Results and Discussion

For our dataset 1 which is most extremely imbalanced with only 0.38% positive class, The f1-score of positive class is 0.60. As we increased the size of positive classes, the f1-score of positive class also improved. The f1-score is 0.90 and 0.94 respectively for dataset 2 and dataset 3. From the table III, we can find that precision and recall improve as well. A model developed for highly imbalanced dataset performs better for more balanced dataset for a similar task of detecting tweets in medication names is also supported by Weissenbacher et al. [2].

We have performed a competitive analysis on dataset 2 about parameter tuning (see table IV). The three key area of parameter tuning are: batch size, embedding, and batch normalization. With dataset being extremely imbalanced, setting up the right batch size is important. From the table IV, we can see that f1-score decreases when batch size is too low or too high. Secondly, embedding tuning improved the performance. When embedding dimension is too high or too low, f1-score decreases as well. The other

[3] glove.twitter.27B.100d: https://pytorch.org/text/_modules/torchtext/vocab.html

Dataset	Positive-class	F1-score	Precision	Recall	TP/FP/FN
Dataset 1	200	0.60	0.70	0.53	26/11/23
Dataset 2	500	0.90	0.95	0.85	105/5/19
Dataset 3	1000	0.94	0.97	0.91	229/8/23

TABLE III: Evaluation of the model on test data

technique which helped to improve performance, is batch normalization. This is possibly because our dataset is highly imbalanced and batch normalization improves the performance of imbalanced dataset. A recent study [3] also shows this.

Dataset	Batch Size	Embedding	Batch Normalization	F1-score
Dataset 2 (default)	128	glove (100d)	yes	0.90
Dataset 2	128	glove (100d)	no	0.89
Dataset 2	32	glove (100d)	yes	0.86
Dataset 2	64	glove (100d)	yes	0.87
Dataset 2	256	glove (100d)	yes	0.82
Dataset 2	128	glove (200d)	yes	0.82
Dataset 2	128	glove (200d)	no	0.86
Dataset 2	128	glove (50d)	yes	0.84
Dataset 2	128	glove (50d)	no	0.85

TABLE IV: Evaluation of the model on various parameters tuning over dataset 2

D. Error Analysis

From the table III, for dataset 2, there are five false positive cases. we will compare our model with Weissenbacher et al. [2]'s eight error category for false positive case for their Kusuri model of similar task: Medical topic, Weighted words/patterns, Ambiguous name, Food topic, Insufficient context, Cosmetic topic, Unknown, Error annotation.

These five cases are listed on table V. If we look through the cases, the number five contains medication name such as 'Codeine'. This was marked as negative class by the annotators. Possibly this is an annotation error. In the other cases, tweets contain words such as food terms (crips, carbs), words often associated with medical topics (meds, cuts, relaxer), ambiguous terms (Prenatal). As medical tweets often describe symptoms and associated with generic medical terms (such as cough, flu, doctor, vitamin, meds, etc), it often confuses the classifier.

Number	Tweet
1	The after fact of having a C section Lawd when the meds ware off I actually feel cut open
2	I need crisps and carbs
3	I want a relaxer... and a blunt cut...
4	My child is apparently auditioning for So You Think You Can Dance: Prenatal Edition. I think he's got a real shot at the title.
5	I'm a suckah For Codeine

TABLE V: False positive cases of dataset 2

From the table III, for dataset 2, there are nineteen false negative cases. we will compare our model with Weissenbacher et al. [2]'s six error category for false negative case: Ambiguous name, Drug not/rarely seen, Generic terms, Nonmedical topic, Short tweets, Error annotation.

These nineteen cases are listed on table VI. If we look through the cases, there are ambiguous name in six (12, 13, 14, 15, 17, 19) cases, Drug not/rarely seen in six (1, 2, 3, 5, 9, 10) cases, Generic terms in three (4, 6, 18) cases, Nonmedical topic in one (16) case, and Error annotation in three (7, 8, 11) cases. The cause of false negative mostly mirrors to false positives.

Number	Tweet
1	@user thanks, I'm 16wks & still taking diclectin... so eventually I'll feel wonderful :)
2	Gaviscon is my new best friend
3	But I do carry a Flonase spray in my purse faithfully bc I have the worst sinus/respiratory infections which cause me to be dizzy
4	@user Try generic (Wal-Mart brand) Prilosec OTC gel capsules. It's the ONLY thing that has helped with my pregnancy
5	@user Hey! He is sleep... still has a Lil anesthesia... but had 2 bottles of Pedialyte and hasn't thrown up... might be home today
6	So I'm watching this commercial about Symbicort inhaler, which I use because I have asthma, and let me tell yall
7	@user @user Got my Tdap at 37 weeks... Hope it wasn't too late. Article states 27-36 weeks.
8	Last time I had a sore arm from a vaccine was when I got my meningitis vaccine
9	A whole 24hrs of me just being itchy af and its nothing I can do about it. Benadryl don't fuckin work g
10	I'm so uncomfortable.. Tylenol3 is my best friend tonight. Gonna snuggle my hubby & crash.
11	Staying in the hospital tonight.Diagnosed with preeclampsia.Was given shotsto make her lungs grow faster, and hoping she stays baking longer
12	My 4th Rhogam shot hurt just as bad as my 1st one. O negative blood is a blessing & a curse. [...]
13	@user how was I supposed to know? It's one full bottle and a few pills in the other. Gotta find them. You can have them.
14	I hate the sleepy side effect when your eggo is preggo or maybe it these vitamins.never wear panties ever but I... http://t.co/vQnof2HBx3
15	Last min oxys so i can sleep... juju knocked... and rell at home cleaning lol
16	Stephanie sounds like she needs some decongestant! [...]
17	@user I am. We are going through 14pts a week & I'm pretty much the only person who has any. And orange Rennies. It's just blurgh.
18	My diet consists of adderall for breakfast, beer for dinner, & cake pops for bed time snack. So I should be skinny again soon.
19	prenatals have done wonders to my hair

TABLE VI: False negative cases of dataset 2

IV. Limitations and Future Work

One of key area, we can improve in our dataset, is preprocessing the hashtags. Usually, the hashtags are combinations of multiple words. But we had not considered this when preprocessing out data. Another major drawback of our system is recognizing medication and dietary names. We have used pre-compiled medications list. The other issue, we had faced, the model was predicting correctly that there's a medication name but the medication terms are abstract - like 'birth control pill', 'meds', etc.

Our future work will involve performance over extremely balanced dataset where positive classes are 0.2-0.3%. And secondly, we will work on to improve the methodology on how to detect medication and dietary names. As the dataset is highly imbalanced, especially, when there are 0.3% positive classes, performance can be improved by upsampling minority class, by downsampling majority class, and adding weight when sampling [4]. Secondly, We can improve the detection of the medication names by using name-entity recognition (NER).

V. Conclusion

In this paper, we have presented a model of learning classifier to identify tweets mentioning medication names. The model was able to achieve an f1-score of 0.94 on an imbalanced dataset with 1.88% positive class. However, it achieved less success for even more imbalanced dataset with 0.38% positive class. The code, dataset, and the model used for these experiments are publicly available [5].

VI. Acknowledgment

The authors acknowledge the help of Mr. Davy Weissenbacher for kindly sharing the dataset.

[4]https://www.tensorflow.org/tutorials/structured_data/imbalanced_data

[5]https://github.com/imranraad07/cmsc-516/tree/main/my_project

REFERENCES

[1] N. Alvaro, Y. Miyao, and N. Collier, "Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations," *JMIR public health and surveillance*, vol. 3, no. 2, p. e24, 2017.

[2] D. Weissenbacher, A. Sarker, A. Klein, K. O'Connor, A. Magge, and G. Gonzalez-Hernandez, "Deep neural networks ensemble for detecting medication mentions in tweets," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1618–1626, 2019.

[3] V. Kocaman, O. M. Shir, and T. Bäck, "Improving model accuracy for imbalanced image classification tasks by adding a final batch normalization layer: An empirical study," *arXiv preprint arXiv:2011.06319*, 2020.