CSE 635: Social Media Mining for Health Monitoring

Tanush Tripathi
Computer Science and Engineering
University at Buffalo
tanushtr@buffalo.edu

Sarveshwar Singhal
Computer Science and Engineering
University at Buffalo
sarveshw@buffalo.edu

Abstract

In this phase of the project we present the final approach to identify tweets which contain an adverse drug reaction (ADR), extract this ADR if present and map it to MEDRA preferred term, classifying Covid-19 tweets containing symptoms and Classifying self-reporting Covid-19 cases using various NLP and ML models. As a part of this Phase we are evaluating the final results for each of the tasks by means of the evaluation metrics such as **F-1 Score,micro F-1 Score, Strict and Relaxed F-1 Score, Precision and Recall**.

1 Introduction

Social media is a popular medium for the public to voice their opinions and thoughts on various health-related topics. Due to the wealth of data available, researchers have been analyzing social media data for health monitoring and surveillance. However, social media mining for health issues is fraught with many linguistic variations and semantic complexities in terms of the various ways people express medication-related concepts and outcomes. The project requires processing imbalanced, noisy, real-world, and substantially creative language expressions from social media to extract and classify mentions of adverse drug reactions (ADRs) in tweets. There are 4 tasks involved in this project:-

Task 1: - Binary Classification of tweets into two classes , one that report adverse effects and the another that doesn't.

Task 2:- Automatic extraction of the span text containing an adverse effect of medication from tweets that report an Adverse Effect (AE)

Task 3:- Classification of COVID 19 tweets containing symptoms into 3 classes, self report , non-personal report, literature and news mentions .

Task 4:- Binary classification of the tweets into two classes, one that self report covid 19 symptoms and the other that does not self report covid 19 symptoms.

2 Literature Survey

The SMM4H Shared task is organized every year with improved or new datasets. We referred to few papers which explained the approaches taken in previous years as well as referred the papers for some of the advanced transformer models we have as the part of this project

2.1 Reference paper on Covid Twitter BERT

The paper defines the need of a Covid Twitter BERT (CT-BERT), a transformer-based model which is pretrained on large corpus of twitter messages on Covid-19. The model was used for Task 3 and 4 for our Milestone 3 which outperformed all the other models on the provided training and validation dataset. This model was Trained on the corpus of 160M tweets about Covid-19 collected from crowdbreaks platform during the period from January 12 to April 16, 2020. Crowdbreaks uses the Twitter filter stream

API to listen to a set of COVID-19-related keywords in the English language. The tweets were then preprocessed and then converted into a set of tokens from a 3000-word vocabulary.

2.2 Reference paper on HealthCare NER Models Using Language Model Processing

The paper provides the approach to extracting structured information from Healthcare data (For example, Data used to study Adverse Drug Reactions), The blank spacy NER model discussed in this paper was used for Task 2 of our Milestone 2 and 3, which gave better results, when compared to other NER models such as Flair and Core NLP. The spacy library contains an EntityRecognizer component, which is a transition-based named entity recognition component. The that identifies non-overlapping labelled spans of tokens.

3 Model Architecture

3.1 Baseline Model

Task 1 As a part of our baseline model, a logistic regression model was used to train the tweets to classify them into an ADR tweet or a non-ADR tweet. We used the Tf-idf Vectorizer to convert the tweets into a matrix of tf-idf values. Logistic Regression, along with the Tf-idf Vectorizer, gave us the best result out of the other results, which included Count vectorizer with Support Vector Classifier.

Task 2 For this task we would perform the preprocessing done in the earlier task. Also, we need to make sure that every token is tagged as per the endpoints given in the training data. The data was combined in a way that all the ADR's for a tweet are stated together. The data was then combined with the starting and ending tag so as to make it ready to be ingested by the NER tagger. The Spacy library was used to train the NER tagger with the training data. For the tweets containing multiple ADR's, the biggest span was considered. After training with the NER Tagger, a logistic regression model, with count vectorizer was used to train the obtained ADR Spans from NER, with the meddra terms and then then meddra terms were evaluated based on the extracted ADR data from the tweets.

Task 3 As a part of our baseline model, a multinomial naive byes classifier was used to train the tweets to classify them into self-reports, non-personal reports, and literature/news mentions. We used the Tf-Idf Vectorizer to convert the tweets into a matrix of Tf-Idf features. Multinomial Naive Byes, along with the TfIdfVectorizer, gave us the best results out of the other results, which included doc2vec with Logistic Regression and doc2vec with Support Vector Machines.

Task 4 Our Baseline model used Logistic Regression to train the tweets to classify them into self-report or non-self-reporting tweets. We have used the Count Vectorizer to convert the tweets into a matrix of token counts. Logistic Regression, along with CountVectorizer, gave us the best results out of the other results, which included Support Vector Machines with Count Vectorizer, Logistic Regression with TfIdf Vectorizer and Support Vector Machines with TfIdf Vectorizer.

3.2 Final Model

Task 1 Our final model was that of using a RoBERTa model with the down sampled/oversampled data, which is a robustly optimized BERT pretraining model that carefully measures the impact of many key hyperparameters and training data size. But before we finalized this model, we also tried training our data on various BERT models such as Basic BERT and BERTweet. We started with the Basic BERT, which is a pretrained model on English language using masked language modelling (MLM).

Task 2 In the Final Model, we started with the spacy's NER model and trained our dataset in order too obtain ADR Extracts from the tweets, like the baseline models. After we computed the extracted ADR Spans for both train and validation dataset using the spacy's ner tagger, we used Sentence Transformer, to vectorize the ADR Spans obtained from NER tagging and the corresponding meddra terms. The reason behind using a sentence transformer is that it is trained on a large number of language prediction tasks that require modelling of meaning of word sequences rather than the individual words. Also, we used cosine similarity to determine the closest possible match between the extracted ADR span and the meddra term vector. Using MiniLM Sentence Transformer, we obtained the best F-1 score compared to all other Sentence Transformers.

Task 3 Our final model was that of using SciBERT, a pretrained language model on Scientific Text to address the lack of high-quality, large scale labeled scientific data. But before finalizing this model, we also tried training our data on various models such as Base BERT (which is a pretrained model on English language using masked language modelling (MLM)), Large BERT(Similar to BERT except that it has 24 encoder layers, double than the former one), BERTweet (which is a large-scale model pre-trained for English tweets) and Covid-Twitter-BERT (which was trained on large corpus of twitter messages on Covid-19 during the period from January 12 to April 16, 2020).

Task 4 Our final model was that of using Covid-Twitter-BERT, a transformer-based model which is pretrained on large corpus of twitter messages on Covid-19. We also tried training our dataset on models such as Base BERT (which is a pretrained model on English language using masked language modelling (MLM)), Large BERT (Similar to BERT except that it has 24 encoder layers, double than the former one) and SciBERT (pretrained language model on Scientific Text).

4 Results

Task 1

Model	**F-1 Score**
Logistic Regression with TfIdf Vectorizer	0.29
Support Vector Machines with Count Vectorizer	0.19
BERT Model - config1(entire data)	0.22
BERT Model - config2(downsampled/oversampled data)	0.32
BERTWEET Model - config1(entire data)	0.20
BERTWEET Model - config2(downsampled/oversampled data)	0.41
ROBERTA Model - config2(downsampled/oversampled data)	0.42

Table 1: Metrics for Task-1

Task 2

Model	Strict F-1	Relaxed F-1	Precision	Recall
Spacy with Logistic Regression	0.20	0.17	0.20	0.21
Spacy with Sentence Transformers				
bert-large-nli-stsb-mean-tokens	0.29	0.35	0.53	0.29
all-mpnet-base-v2	0.26	0.31	0.48	0.26
all-distilroberta-v1	0.26	0.30	0.49	0.26
all-MiniLM-L12-v2	0.31	0.37	0.57	0.31

Task 3

Model	**F-1 Score**
Multinomial Naive Byes with TfIdf Vectorizer	0.86
Logistic Regression with doc2vec	0.61
Support Vector Machines with doc2vec	0.57
BERT-base-uncased	0.95
BERT-large-uncased	0.94
BERTweet (bertweet-covid19-base-uncased)	0.96
Covid-Twitter-BERT	0.97
SCI-BERT	0.97

Table 2: Metrics for Task-3

Task 4

Model	F-1 Score
Logistic Regression with Count Vectorizer	0.40
Support Vector Machines with Count Vectorizer	0.27
Logistic Regression with TfIdf Vectorizer	0.30
Support Vector Machines with TfIdf Vectorizer	0.28
BERT-base-uncased	0.28
BERT-large-uncased	0.18
Covid-Twitter-BERT	0.34
SCI-BERT	0.53

5 Discussion adn Error Anallysis

1.Neutral data: There were certain tweets which were neutral in the dataset but were categorized as negative. Tweet like [‘tweet_id’: 344266386467606000, ‘tweet’: “depression hurts, cymbalta can help”] This tweet says about a problem and it’s cure. It doesn’t mention about whether that cure has any ADR or not.

2. We felt few of tweets were misclassified as negative, instead of positive. Tweet like [‘tweet_id’: 326594278472171000, ‘tweet’: “@redicine the lamotrigine and sjs just made chaos more vengeful and sadistic.”] here the term ‘lamotrigine’ is an antidepressant, but the tweet is saying user felt ‘sadistic’, which is exactly opposite of what drug is suppose to do.

3. In the dataset we found some sarcastic tweets. These kind of tweets creates ambiguity in the model. Twee like [‘tweet$_id'$: $342038595890196000, ‘tweet'$: $\backslash thankgodforhumira\ddot{Y}Œwonderdrugcrohnsdisease$”]; $Herewedon'tknowwhethertheuserwantstoreportADRornot.$

6 Conclusion

Although we were able to come near the actual values by the students in the SMM4H Conference, we have still a lot of work to do. As the dataset was heavily biased we need to come up with different approaches which can help us to reduce the biasedness in the data. We plan to do more research on some more approaches like T5 etc. to further fine-tune our models and achieve a higher performance in terms of F-1 Scores as well as to work with a larger collection of tweets.

7 Work Distribution

Distribution of work Tanush	Distribution of Work Sarveshwar
Task 2 and 3 Baseline and Final Code	Task 1 and 4 Baseline and Final Code

Distribution of Work Tanush

Implemented the Baseline and Final models for the Tasks 2 and 3 , which included,

1. Running NER Tagging for Task2 , accompanied by logistic regression and sentence transformers for baseline and final models respectively.

2. Training Multinomial Naive Byes, Logistic Regression and Support Vector Classifier models with tf-idf and doc2vec for baseline analysis, and running various BERT Models for final Analysis for Task3

Distribution of Work Sarveshwar

Implemented the Baseline and Final models for the Tasks 1 and 4 , which included,

1. Training a logistic Regression model d=and Support Vector Classifier with Tf-idf vectorizer and count vectorizer and different BERT Models

2. Training Logistic Regression and Support vector classifier with tf-idf and count vectorizer for baseline analysis, along with training data with various BERT Models for final Analysis, For Task4

8 Bibliography

1.https://huggingface.co/lordtt13/COVID-SciBERT
2.https://huggingface.co/bert-large-uncased
3. https://spacy.io/api/entityrecognizer
4. https://arxiv.org/abs/2005.07503
5. https://arxiv.org/ftp/arxiv/papers/1910/1910.11241.pdf