Analysis of Tweets for Social Media Health Applications

by

Shubham Gondane

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2019 by the
Graduate Supervisory Committee:

Chitta Baral, Chair
Saadat Anwar
Murthy Devarakonda

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Social networking sites like Twitter have provided people a platform to connect with each other, to discuss and share information and news or to entertain themselves. As the number of users continues to grow there has been explosive growth in the data generated by these users. Such a vast data source has provided researchers a way to study and monitor public health. Accurately analyzing tweets is a difficult task mainly because of their short length, the inventive spellings and creative language expressions. Instead of focusing at the topic level, identifying tweets that have personal health experience mentions would be more helpful to researchers, governments and other organizations. Another important limitation in the current systems for social media health applications is the use of a disease-specific model and dataset to study a particular disease. Identifying adverse drug reactions is an important part of the drug development process. Detecting and extracting adverse drug mentions in tweets can supplement the list of adverse drug reactions that result from the drug trials and can help in the improvement of the drugs.

This thesis aims to address these two challenges and proposes three systems. A generalizable system to identify personal health experience mentions across different disease domains, a system for automatic classifications of adverse effects mentions in tweets and a system to extract adverse drug mentions from tweets. The proposed systems use the transfer learning from language models to achieve notable scores on Social Media Mining for Health Applications(SMM4H) 2019 (Weissenbacher et al. 2019) shared tasks.

To Mom and Dad

ACKNOWLEDGMENTS

I would like to thank the people involved in this project for their continuous support and guidance. First and foremost, my thesis advisor, Dr. Chitta Baral, who has guided me at every step in this project with his expertise and knowledge. Thanks to my committee members Dr. Saadat Anwar and Dr. Murthy Devarakonda for their help on this thesis. Thanks to Arizona State University for providing me an opportunity to work with such great minds and cutting-edge technology. Thanks to my colleagues in Dr. Baral's lab, especially Pratyay, Kuntal, Kuldeep, and Swaroop for their help with preparing my presentation. A special thanks to Sam who has been a great help in brainstorming different ideas and providing last minute inputs. Last but not least, thanks to my parents, and numerous friends who have endured this long journey with me for with their constant support and love.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION AND MOTIVATION

1.1 Introduction

The latest Pew Research Report [1], states that nearly half of adults worldwide and two-thirds of all American adults (72%) use social networking. Social media mining is the process of extracting and analyzing patterns from user data available online. The information available online is often used for marketing campaigns, advertisement, capturing consumer feedback, keeping track of competing products. There is another critical use of the social media data as highlighted by the Pew Research Report which states that of the total users, 26% have discussed health information, and, of those, 30% changed behavior based on this information, and 42% discussed current medical conditions. Around 500 million tweets are posted every day, according to recent estimates. Such a massive data source that can be used to develop numerous public health applications. Recent advances in machine learning and Natural language processing make it an exciting opportunity to work on social media health mining problems.

Studies done in the past have focused on identifying different health topics on Twitter. Various health issues like Allergies, Aches/Pains, Cancer, Obesity, Flu, etc. have been discovered to be discussed on Twitter (Paul and Dredze 2011). Extensive work has been done in investigating about specific diseases like tracking the spread of flu epidemics (Lamb, Paul, and Dredze 2013), cancer analysis (Lee, Agrawal, and

[1] https://www.pewinternet.org/fact-sheet/social-media/

Choudhary 2013),and depression prediction (De Choudhury et al. 2013). Such studies have had a significant impact on public health, medical resource allocation, health policy, and education (Neiger et al. 2013).

1.1.1 Generalizable Systems

The studies have proved the effectiveness of analyzing and monitoring public health. There is still a long way to go towards building more robust systems that can automatically identify a health mention in tweets. A lot of traditional systems underperform, especially at data that is usually imbalanced, noisy and consist of creative language expressions. Often to build a classification model to track a particular disease or health concern, it becomes necessary to use a health concern specific data set. Such a model only works for that particular health concern and performs poorly for other health issues. Building a general model to classify tweets across various health concerns would be of great use to the research community and allow using existing tools and techniques to solve new problems. *Another type of generalization is in building model that work well on future data as well. Previous work has shown that classification performance usually varies across different time intervals* (Huang and Paul 2018).

1.1.2 Adverse Drug Reaction Detection and Extraction

Adverse Drug Reaction(ADR)

An Adverse Drug Reaction is any injury that is caused by medication intake. It can range from simpler side effects like headaches, drowsiness to more complicated ones like heart issues or skin rashes. An adverse effect mention is any injury that occurs during the medication intake but it's cause is not yet attributed to the drug. ADR is a specific case of an adverse effect mention.

Every year the pharmaceutical industry spends millions of dollars developing new drugs. Identifying adverse drug reactions(ADR) is extremely important part of the drug development process. Pharmacovigilance is a set of tasks relating to the collection, detection, assessment, monitoring, and prevention of adverse effects with pharmaceutical products. During its development process drugs are subject to various trials which can often reveal any ADR resulting from the use of the drug. However, the ADR data collected in such trials is limited, centralized and often not exhaustive. Once the drug hits the market, it becomes difficult to collect ADR data from numerous sources and report the health professionals. It is one of the most significant and expensive public health problem. To solve this problem the pharmacovigilance has turned to social media in order to monitor the adverse drug reactions that develop outside the clinical trial and tests. The number of people sharing details about their medication intake has increased in recent years. Identification and extraction of adverse effect mentions in tweets has proved to be a challenging task because of the excessive use of hyperbole, sarcasm, and ambiguous language used by people online.

1.2 Social Media Mining for Health Applications Shared Tasks

The Social Media Mining for Health Applications (SMM4H) Workshop 2019 (Weissenbacher et al. 2019) proposed four tasks to address these challenges. These thesis addresses 2 of those tasks.

1.2.1 Generalizable System To Identify Personal Health Mentions

The challenge is a binary classification task to classify whether a given tweet contains a personal health mention as opposed to a general comment, suggestions, or unrelated mention of the health issue.

Positive example:

Personal mention: *"However, I worked hard and ran for Tokyo Mayer Election Campaign in January through February, 2014, without publicizing the cancer"*

Negative examples:

General comment: *"A Month Before a Heart Attack, Your Body Will Warn You With These 8 Signals"*

Unrelated mention: *"Now I can have cancer on my wall for all to see <3"*

Previous work has shown that people often use words like "heart attack" or "cancer" for emphasis or convey sarcasm. Therefore, the number of such tweets drastically outnumber the personal mention tweets. Combining this with the short length of tweets and the presence of typos, creative spellings, and misspellings make it challenging.

The training data in this task has tweets of one health domain - influenza across two contexts flu vaccination or flu infection. The test data will contain tweets from the same domain across different context and also a completely different health domain in some context.

1.2.2 Automatic Classification of Adverse Effect Mentions in Tweets

The task is to build a system that can distinguish between a tweet reporting an adverse effect mention from those that do not.

Adverse effect mentions are unwanted or harmful reactions resulting from correct medical use.

For example:

Presence of adverse effect: *Not that anyone noticed, but my #**ambienwithdrawl** only lasted a few days. Why? Because I got another scrip. I need it while I'm on Levaquin.*

Absence of adverse effect: *@C4Dispatches Eeeeek. Just chucked my Victoza in the bin. I will take my chances with the diabetes #**diabetes***

1.2.3 Extraction of Adverse Effect Mentions

The task is to build a system that can identify and extract adverse drug reaction (ADR) spans from a tweet. ADRs are multi-token, descriptive, expressions so this is essentially a Named Entity Recognition (NER) challenge.

For example:

another night of 'light' sleep. feel low. subdued. sad. #venlafaxine #day12

ADR extraction: feel low

Span: 32 40

1.3 Related Work

There has been limited research done towards building general models to classify different diseases. (Paul and Dredze 2011) use an LDA topic model-based system to discover mentions of over a dozen ailments. They further incorporate prior knowledge into this model and apply it to tasks like tracking illnesses, measuring behavior risk.

(Prieto et al. 2014) proposed a two-step process to detect the health mentions in social text data. The first step is to collect the tweets using keywords and regular expressions, and the second step is to use a high-precision classifier – in this case, by using a correlation-based feature extraction method. The system was designed for the analysis of flu, depression, pregnancy, and eating disorders.

(Yin et al. 2015) reported using a dataset of tweets across 34 health topics and investigated the accuracy of the classifiers trained over multiple diseases and tested on new diseases. The authors conclude that training a classifier on four diseases: cancer, depression, hypertension, and leukemia can lead to a general health classifier with 77% precision using standard SVM classifiers.

(Karisani and Agichtein 2018) developed a system called as WESPAD that combines lexical, syntactic, word embedding-based, and context-based features. The authors report that the system can generalize from a few examples by automatically distorting the word embedding space to most effectively detect the true health mentions.

(Kiritchenko et al. 2018) developed support vector machine classifiers using a variety of surface-form, sentiment, and domain-specific features. To reduce of class imbalance they use an under-sampling technique.

(Hakala et al. 2017) use an ensemble of neural networks with features generated by word and character-level convolutional neural network channels and a condensed weighted bag-of-words representation

(Xherija 2018) use a variant of the Message-level Sentiment Analysis (MSA) a word- level stacked bidirectional Long Short-Term Memory (LSTM) network.

(Shen et al. 2018) used a neural approach with hierarchical tweet representation and multi-head self-attention. Their system uses a three module approach where the first module learns the word representations from the character embeddings, next the module learns the tweet representation from the words and finally a classification module. This work has been extended by (Ge et al. 2019) where they utilize additional features like word2vec Twitter(Godin and Vandersmissen 2015), POS tags and sentiment lexicon along with embeddings from a language model ELMo(Peters et al. 2018).

(Shen et al. 2018) combine different word level embeddings and use a multi channel Convolutional Neural Networks to identify adverse drug reaction information. The multiple channels help to learn features from character level embeddings of the words.

(Nikfarjam et al. 2015) introduce ADRMine, a machine learning based extraction approach that uses conditional random fields (CRF) using variety of features like ADR lexicon, Part Of Speech(POS), negation and context based features.

(Gupta et al. 2018) introduced a joint multi-task learning method that uses adverse drug event detection as an auxiliary task to improve performance on the ADR extraction task. They generate weak supervision dataset for the auxiliary task using a large pool of unsupervised dataset.

(Sarker and Gonzalez 2015) use a large set of features based on semantic properties such as sentiment, polarity and topic. They also combine training of different corpora to enhance the classifier performance.

1.4 An Overview to the Approach

Generalizable system for personal health mentions:

We use the feature representations generated by the BioBERT model fine-tuned on the training data. In this approach the tweets are first preprocessed where in, the underlying tweet text is cleaned and some twitter specific features are extracted. The next step is to replace(term generalization) all the mentions of a disease in the tweet.

Our system uses the the BioBERT model that is fine-tuned on original tweets. This fine-tuned model is then used to extract representation for both the term generalized tweets and original tweets. The feature representation are the weights of the last 4 layers of the BioBERT model for the CLS token is concatenated together to form a 3072 dimensional vector. The feature representation for term generalized tweets and original tweets are concatenated together. These are input to a fully connected dense neural network with two hidden layers.

In order to enhance the performance of the system we create an ensemble model where we combine the weights of 10 different models based on their performance on the development data set.

Automatic classifications of adverse effects mentions in tweets: We have used the feature representation generated by the BioBERT model fine-tuned on the training data. In this approach the tweets are just preprocessed and the feature representation for the tweets are extracted from the fine-tuned BioBERT model. The representation is then run through a neural network with two hidden layers and a softmax output layer.

In order to improve the performance of the system we use additional datasets like the CADEC corpus(Karimi et al. 2015). We also collected around 750K tweets related to ADRs to pretrain the BioBERT model.

Extraction of adverse effect mentions in tweets: We use the BioBERT model with the Named Entity Extraction task and modify it by adding a Conditional Random Field layer on top of the model to get more accurate predictions. The system can be improved by using external datasets like SIDER(Kuhn et al. 2015), CADEC(Karimi et al. 2015) along with pretrainng the BioBERT model on around 750K tweets.

The rest of the Thesis is organized as follows. We go over the recent progress and the background behind the language models. We describe the datasets, model descriptions, evaluation criteria and results for each of the tasks in subsequent chapters.

Chapter 2

LANGUAGE MODELS

Transfer learning is a technique used in areas like computer vision and more recently NLP where the model that has been trained on a particular task is used for a different but related task. This process is often used to jump-start the development of a new model by fine-tuning the pretrained model. Typically the weights learned by the pre-trained model are used to initialize the new model and the weights are re-learned. The main benefit of transfer learning is that it can speed up the model development process considerably. Another advantage is that the pre-trained models can be used on a smaller dataset to achieve good results. Recent progress in NLP has come from the use of transfer learning. Language models like BERT (Devlin et al. 2018), ELMo (Peters et al. 2018) and OpenAI-GPT (Radford et al. 2019) have had considerable success in wide variety of NLP tasks.

2.1 BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers is a language representation model that is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. At the time of its release BERT obtained state-of-the-art results on eleven NLP tasks. Pretrained representations are often used in a downstream tasks with the help of two techniques - *Fine-tuning and feature embeddings.*

We have experimented with both the above mentioned techniques in the aim of obtaining better results. The authors have released two versions of BERT model - base and large. We have used the BERT-Base model in the experiments which has the following attributes:

2.1.1 Transformer Specifications

BERT's architecture is a multi-layer bidirectional Transformer encoder. In the BERT-Base, model there are 12 transformer layers , 768-hidden states, 12 attention heads resulting in overall 110M parameters.

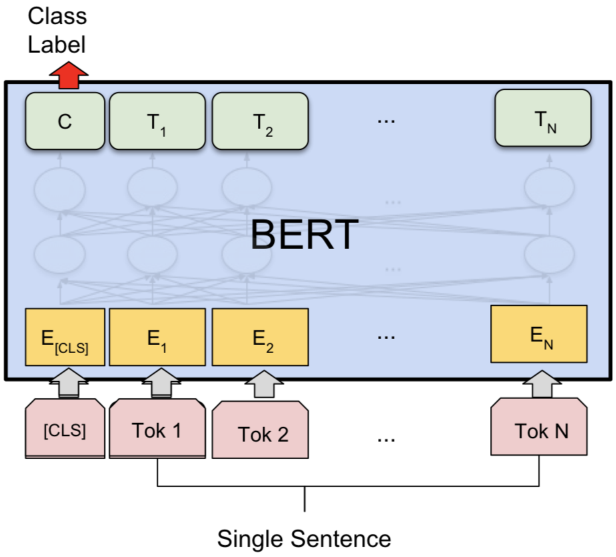


Figure 1: How to use BERT for classification

Source: (Devlin et al. 2018)

2.1.2 Tokenization

The BERT model learns the contextual representations using both the left and the right contextual information starting from the very first layer of the network. To achieve the deep bidirectionality in the model BERT uses an approach that masks out 15 percent of the words in the input and runs the entire sequence through a deep bidirectional Transformer encoder, and then predicts only the masked words. BERT uses its own tokenization to achieve the masking of words and it also adds the following special tokens.

CLS Token: A classification token which is normally used in conjunction with a softmax layer for classification tasks. For anything else, it can be safely ignored.

SEP Token: A sequence delimiter token which was used at pre-training for sequence-pair tasks (i.e. Next sentence prediction). Must be used when sequence pair tasks are required. When a single sequence is used it is just appended at the end.

After the model is fine-tuned, the features are extracted which are nothing but the contextual representation generated for the individual tokens and the CLS and SEP tokens. There are a number of ways to use the hidden layer representations - using the final hidden layers, concatenating last few hidden layers, summing up last few hidden layers and so on. The authors state that using a concatenation of last 4 hidden layer representation gives the best result on the various tasks. We have used this very same combination in our experiments.

2.2 BioBERT

BioBERT[2] which stands for Bidirectional Encoder Representations from Transformers for Biomedical Text Mining is a domain specific language representation model based on the BERT architecture. Though the BERT model has two versions base and large the BioBERT model only uses the BERT-Base model. So the attributes of the BioBERT model are same as that of BERT.

2.3 RoBERTa

A Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al. 2019) builds on BERT's language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples. RoBERTa modifies key hyperparameters in BERT, including removing BERT's next-sentence pretraining objective, and training with much larger mini-batches and learning rates. This allows RoBERTa to improve on the masked language modeling objective compared with BERT and leads to better downstream task performance. RoBERTa has achieved better results on the GLUE Benchmark compared to BERT We use RoBERTa-base in our models.

[2] BioBERT version 1.0

2.4 Fine-tuning

Fine-tuning is one way of taking advantage of pre-trained representation. The fine-tuning process remains same for the BERT and BioBERT models with the exception of the pre-trained weights[3,4] used in those models.

Fine-tuning is done end-to-end by providing the data and the labels. The authors have provided a set of data processors to handle the input to the model. We have used the CoLA processor and converted our data to the CoLA (GLUE version (Wang et al. 2018)) dataset format. The dataset was converted into the following format:

Tweet id | Label | * | Tweet text

The fine-tuning process has very high variance in dev set accuracy for smaller datasets. So for our experiments we ran the fine-tuning process for 5 iterations every time and the best performing model is used further.

2.5 Feature Extraction

Another way to using the pre-trained representations in a downstream task is to extract the fixed contextual representations from the pre-trained model and use them in another model. These representations are the weights of the neurons in the hidden layers. Here we have used feature extraction after fine-tuning the model. The reason behind this order is that since tweets are quite different from the usual English

[3] BERT pre-trained weights used: https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

[4] BioBERT pre-trained weights used: https://github.com/naver/biobert-pretrained/releases/download/v1.0-pubmed/biobert_pubmed.tar.gz

language text the models were trained on it makes sense to fine-tune the pre-trained model so the weights get updated according to the new data.

There are two ways to use word embeddings in a neural network. The first is to learn the word embeddings for a particular problem. This approach requires a large amount of training data. The second approach is to use a pre-trained word embeddings like word2vec and GloVe(Pennington, Socher, and Manning 2014). These models have a vocabulary associated with them and an embedding matrix where each row is a vector representation for a particular word. In case of BERT there is no embedding matrix which can be directly loaded since it takes care of tokenization and initializing the model, so we need to create our own custom embedding matrix to take care of this issue. To achieve this we tokenize the data using the tokenizer provided and create a mapping of the tokens to the token ids. Once we have this mapping we use the fine-tuned model and get the weights of the final layer. So we create a 768-dimensional feature representation for each of the words in the vocabulary, thereby generating our own embedding matrix.

2.6 Pre-training

The performance of the language model largely depends on the corpora on which it was trained on. Since BERT was trained to be a general purpose language model it was trained on Wikipedia and Book Corpus. Since biomedical text often has domain specific proper nouns like (eg. BRCA1, c.248T>C) and terms (e.g., transcriptional, antimicrobial) which are understood mostly by biomedical researchers.(Lee et al. 2019) So general purpose models perform poorly on biomedical NLP tasks. BioBERT improves upon BERT by pretraining on additional biomedical datasets like PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC).

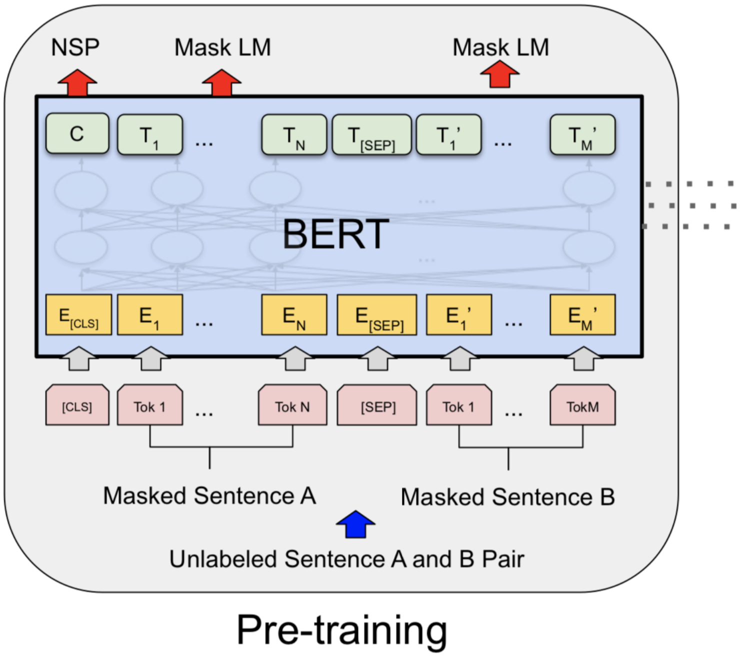


Figure 2: BERT pre-training

Source: (Devlin et al. 2018)

Tweets often consist of novel/creative phrases and misspellings, and frequent use of idiomatic, ambiguous and sarcastic expressions (Weissenbacher et al. 2019). Thus in order to increase performance we have pretrained BioBERT model on a corpus of medical tweets. There are two unsupervised tasks that are used to pre-train the model - **Masked language modeling and Next sentence prediction.**

Masked Language Modeling

Since BERT is deep bidirectional model which allows each word to indirectly "see itself" and the model would trivially predict the target word. So to train bidirectional representations the BERT/BioBERT masks some percentage of the input tokens and then predict those masked tokens.

Next Sentence Prediction

In order to understand the relationship between two sentences which is essential for NLP tasks like natural language inference and question answering, BERT has been trained on the next sentence prediction task.

During pre-training process we have used the masked language modeling task and not the next sentence prediction. The reason behind this is that tweets are often just one sentence long and so the next sentence prediction task does not make sense here. The exact parameter used to pre-train the model are mentioned in the appendix.

Chapter 3

GENERALIZABLE SYSTEM TO IDENTIFY PERSONAL HEALTH MENTIONS IN TWEETS

3.1 Dataset Description

The data set provided by the SMM4H organizers consisted of two data sources across different contexts - flu infection and flu vaccination both in the flu domain. Each data set consisted of Tweet ids and the label indicating the presence of personal health mention. The flu infection data set had 1046 records of flu infection, but around 1023 tweets were available for download. The flu vaccination data set had around 9800 records out of which only 6659 were available for download. The tweets were downloaded using the Twitter API. The combined data set has 7682 tweets in total.

Approximately 54 percent of the data in the flu infection set is labeled as positive, and approximately 30 percent of the data in the flu vaccination set belongs to the positive class.

For example:

Positive: Got my **flu** shot today &amp; my body hurts so bad. Especially my arm

Negative: NJ Getting **Swine Flu Vaccine** Next Week http://ff.im/-8URDd

The test data included tweets from the influenza domain and also tweets from a separate domain, in order to help test the generalizability of the classifiers. Testing data from the influenza domain contains data collected years after the original training data, to test the generalizability of the classifiers to future data.

Data	Context	Tweets
Train set	Flu infection	1023
	Flu vaccination	6659
Test set	Flu and unknown health context	285

Table 1: Dataset Description

3.2 Term Generalization

The challenge in this task is to train a model on one disease domain and test on another, so it is necessary to introduce generalization in the model. One way to achieve this is to mask specific terms like flu or influenza mentions with an AILMENT token. A list of all flu-related terms was created using a pretrained Word2Vec model for Twitter (Godin and Vandersmissen 2015) to find similar terms to flu. The list was expanded using human knowledge and ConceptNet[5] (Speer, Chin, and Havasi 2017). This list of terms was used to replace all the flu mentions in the dataset. We have also experimented with some models that use the tweets without using the generalized term to replace the flu mentions.

[5] www.conceptnet.io

3.3 Preprocessing

The preprocessing library Ekphrasis provides several methods to clean and process social media text (Baziotis, Pelekis, and Doulkeridis 2017). This library is used to preprocess all the tweets.

- All @user mentions were replaced by user token.
- All HTTP URLs were replaced by url token.
- Hashtags were preprocessed by removing the # symbol and keeping the words.
- Emojis, dates, numbers, etc. are removed.
- The text is converted to lowercase.

For example:

Raw tweet:

Calling for medical advice. Verdict already in though: ***Tonsillitis*** *(& getting* ***flu*** *treatment just in case).* ***http://twitpic.com/pk2q0***

Preprocessed tweet without term generalization:

calling for medical advice . verdict already in though : ***tonsillitis*** *(& getting* ***flu*** *treatment just in case) . url*

Preprocessed tweet with term generalization:

calling for medical advice . verdict already in though : ***ailment*** *(& getting* ***ailment*** *treatment just in case) . url*

3.4 Traditional Machine Learning approaches

Traditional feature-based models have found some success in building generalizable systems. We have experimented with such feature-based logistic regression and Support Vector Machines.

3.4.1 Features

Before proceeding with the models and the results obtained, we first explain the different features that are used in our experiments. These features are only used in the logistic regression and SVM models.

3.4.1.1 N-grams

N-grams are used extensively in traditional NLP models. N-grams are set of co-occurring words within a given range. The N-gram range varies according to the application. Since the language used on Twitter is short in length, we have experimented with n-grams in the range 1 to 3 i.e., unigrams, bigrams, and trigrams.

As we are replacing the presence of flu-mentions, we create one n-gram feature set without replacing the flu mentions and another with the flu-mentions replaced. So this way we have both domain dependent and domain-independent set of features.

3.4.1.2 Twitter Specific Features

Tweets have a 280-character limit and given such a small number of characters available users often resort to the use of emojis[6] to convey themselves better. Tweets often have links used to direct the people to different web pages. Use of hashtags allows users to share their post with other people who are interested in the same topic. All these parts of a tweet can be used as features as shown by the authors in their work on classifying adverse drug mentions in tweets (Kiritchenko et al. 2017).

- Number of hashtags
- presence or absence of emojis
- Number of elongated words
- Number of words with all letters in uppercase
- Number of question marks
- Number of exclamation marks

These features are extracted before preprocessing the tweet.

Example: "@hot995: VIDEO: Watch as this girl getting a FLU SHOT goes NUTS! LOL! http://t.co/M73KmdK9lr http://t.co/MlDDDzUxAD"

Feature vector generated: [0, 1, 0, 4, 2,0]

[6] https://en.wikipedia.org/wiki/Emoji

3.4.1.3 Sentiment Scores

For each token in the tweet we use a lexicon containing approximately 155,000 English words associated with a sentiment score between -1 and 1.(Liu, Hu, and Cheng 2005)(Sarker and Gonzalez 2015)

Example words:

Positive words: awed, best, cozy, excite, fervor

Negative words: ache, adverse, anxiously, concern, despair

3.4.2 Logistic Regression Model

Logistic regression is a technique under the supervised learning paradigm that works quite well for various classification tasks. Here we have tried various logistic regression models with different feature sets as follows:

LR Model 0: Baseline Logistic Regression Model with n-grams without preprocessing.

For our baseline model, we experimented with combinations of different n-gram with n from 1 to 5 as features on unprocessed tweets. The best score resulted from a combination of unigrams, bigrams. This is more likely because of the short length of tweets. We decided to use this n-gram range in further models.

LR Model 1: Logistic Regression Model with n-grams with preprocessing.

In this model, we experimented with different combinations of unigrams and bigrams as features on preprocessing tweets without replacing the flu-mentions with generalized term.

LR Model 2: Logistic Regression Model with n-grams with preprocessing and term generalization

This model uses unigrams, bigrams as features on preprocessing tweets with with generalized flu-mentions.

LR Model 3: Logistic Regression Model with top 1200 n-grams with preprocessing and term generalization.

Instead of using the entire vocabulary of n-grams, we decided to experiment with a particular number of features. We used the previous models to find the ideal number of features. Logistic regression model gives weight to the features, the higher the weight the more relevant the feature is for positive class and vice-versa for the negative class.

Top features for positive class:

- getting over
- pain
- fever
- health emergency
- very sick

Top features for negative class:

- url
- my immune
- not got
- feel better
- someone

The graph below represents the f1 score comparison between different number of features.

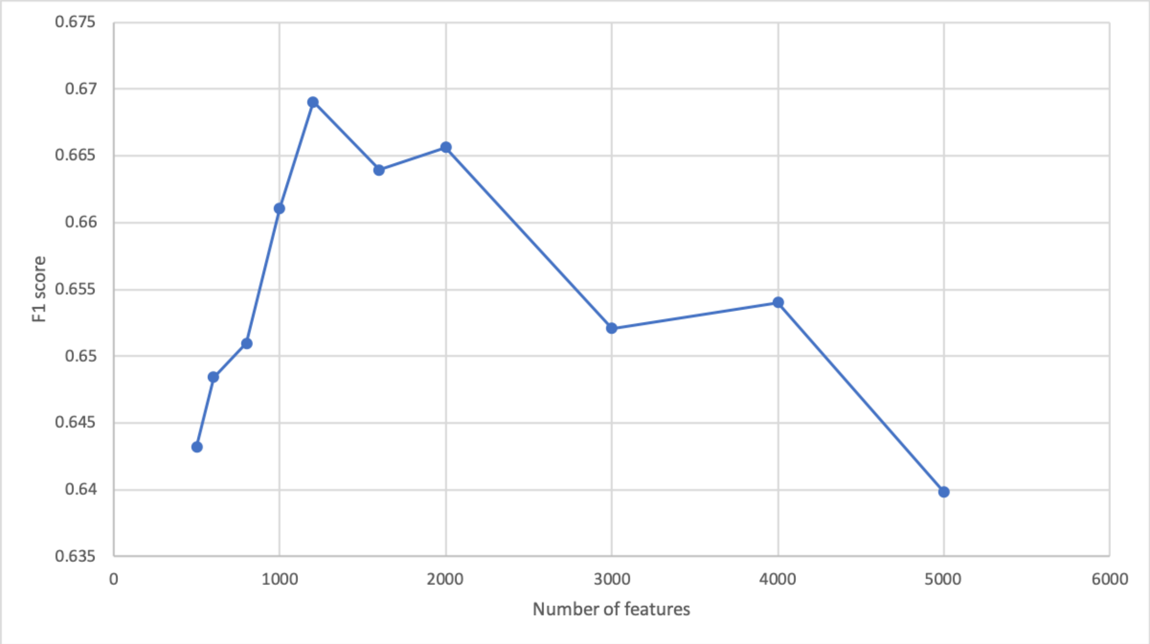


Figure 3: Number of Features vs F1 Score

LR Model 4: Logistic Regression Model with top 1200 n-grams with preprocessing and generalized tokens combined with top 1200 n-grams with preprocessing and original tokens.

This model used n-grams of both generalized and original tweets as features

LR Model 5: Logistic Regression Model with top 1200 n-grams with preprocessing and generalized tokens and Twitter specific features.

This model uses Twitter specific features along with textual features.

LR Model 6: Logistic Regression Model with top 1200 n-grams with preprocessing and generalized tokens and Twitter specific features and Sentiment features.

This model uses sentiment scores along with the previous set of features.

This model uses change phrase features along with the previous set of features.

LR Model 7: Best Logistic regression model with top 1200 n-grams with preprocessing and generalized tokens and selective Twitter specific features.

3.4.2.1 Feature Selection

As we used feature selection for selecting the top 1200 n-grams, we used the same technique to see whether the Twitter-specific features, sentiment scores, and change phrase features were contributing to the classification decision.

Remarkably change phrase feature did not make any contribution to the model as the feature was mainly designed for adverse drug reaction tweets. Also, sentiment scores and some of the Twitter-specific features did not contribute to the classification process. Out of the 7 Twitter specific features, only two features - *number of exclamation marks and number of question marks* had weight more than 0.

Model	F1 score
LR Model 0	0.61
LR Model 1	0.63
LR Model 2	0.66
LR Model 3	0.67
LR Model 4	0.67
LR Model 5	0.67
LR Model 6	0.68
LR Model 7	0.71

Table 2: Logistic Regression Model Scores

3.4.2.2 Training Process

All models follow a similar training process. We used k-fold cross validation with train test split of 80:20.

3.4.3 Support Vector Machine Model

Similar to the various logistic regression models mentioned above we also used an SVM classifier on the different feature configurations as used in the logistic regression section.

SVM Model 1: SVM Model with n-grams with preprocessing

In this model, we experimented with different combinations of unigrams and bigrams as features on preprocessing tweets without replacing the flu-mentions.

SVM Model 2: SVM Model with top 1200 n-grams with preprocessing and term generalization

This model used n-grams of both generalized and original tweets as features

SVM Model 3: SVM Model with top 1200 n-grams with preprocessing and term generalized tokens and Twitter specific features

This model uses Twitter specific features along with textual features.

SVM Model 4: SVM Model with top 1200 n-grams with preprocessing and term generalized tokens and Twitter specific features and Sentiment features

This model uses sentiment scores along with the previous set of features.

SVM Model 5: SVM Model with top 1200 n-grams with preprocessing and term generalized tokens and Twitter specific features and Sentiment features and change phrase features

This model uses change phrase features along with the previous set of features.

Best SVM model with top 1200 n-grams with preprocessing and generalized tokens and Twitter specific features.

While the Logistic regression and SVM models performed well, we decided to experiment with language models like BERT and see how they perform on the text that's present in tweets. In particular we wanted to see if the language structure of the tweets have any effect on the BERT models' performance.

Model	F1 score
SVM Model 1	0.78
SVM Model 2	0.79
SVM Model 3	0.81
SVM Model 4	0.82
SVM Model 5	0.84

Table 3: SVM Model Scores

3.5 Transfer Learning Approach

We tried both the fine-tuning based and feature embeddings approaches. The background for both these tasks is explained in the previous chapter. We describe the various models we experimented with here using these transfer learning approaches here. At the time of writing the organizers of the SMM4H shared task 2019 did not release the test dataset gold labels. All the transfer learning models were evaluated on the system provided by the organizers.

3.5.1 End-to-End Model

TL Model 0: BERT model fine-tuned end-to-end without any preprocessing

This is a baseline model using just raw tweets.

TL Model 1: BERT model fine-tuned end-to-end

This is BERT base model trained on preprocessed tweets with term generalization.

TL Model 2: BioBERT model fine-tuned end-to-end

This is BioBERT base model trained on preprocessed tweets with term generalization.

3.5.2 Fully Connected Neural Network Model

The following three models use concatenated feature representation of the CLS token from the last 4 layers of the fine-tuned BERT/BioBERT model. The concatenation is a 3072-dimensional vector. Figure 2. represents the neural network architecture used in the model. A dropout layer is added between the two hidden layers, and the hyperparameters are tuned accordingly.

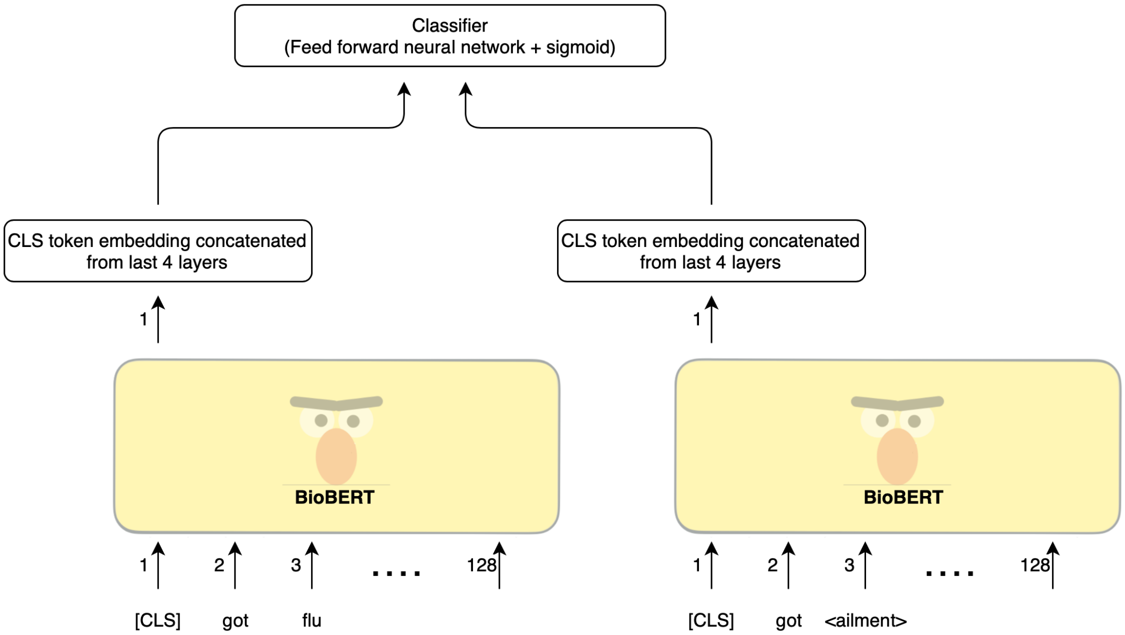

Classifier
(Feed forward neural network + sigmoid)
CLS token embedding concatenated from last 4 layers
CLS token embedding concatenated from last 4 layers
1
1
BioBERT
BioBERT
1
2
3
....
128
1
2
3
....
128
[CLS]
got
flu
[CLS]
got
<ailment>

Figure 4: Generalizable System Model Architecture

Source: Figure adapted from (Alammar 2018)

Model	Accuracy	F1 score	Precision	Recall
TL Model 0	0.82	0.8	0.79	0.82
TL Model 1	0.86	0.85	0.86	085
TL Model 2	0.87	0.85	0.87	0.85
TL Model 3	0.90	0.89	0.94	0.86
TL Model 4	0.91	0.91	0.97	0.85
TL Model 5	**0.93**	**0.92**	**0.97**	**0.88**

Table 4: Transfer Learning Models Scores

Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.84	0.80	0.97	0.68
health concern condition 1	0.92	0.92	0.98	0.86
health concern condition 2	0.69	0.51	0.91	0.35
health concern condition 3	0.80	0.59	1	0.42

Table 5: Final Scores on the Test Set for Fully Connected Neural Network.

- **TL Model 3: Neural Network with BioBERT original embeddings**
- **TL Model 4: Neural Network with BioBERT generalized embeddings**
- **TL Model 5: Neural Network with BioBERT original embeddings and generalized embeddings**

The TL model 5 was our official submission to the shared task 4 challenge. The results prove that the transfer learning approach proves to be promising for building generalizable system for identifying personal health mention. Table 5 describes the evaluation scores of the system.

3.5.3 Improvement: Ensemble Model

Each time a neural network model is trained the model we get a different set of weights. Using an ensemble models helps to leverage these different set of weights by combining them in some way. One way to combine the weights is to take average of all weights and the reason that model averaging works is that different models will usually not make all the same errors on the test set (Bengio, Courville, and Goodfellow 2016). We improve upon our submission by using an ensemble of the BioBERT model with generalized and original embeddings. We used a total of 10 models. During fine-tuning process we make sure to choose the model which gives the best accuracy on the development set, then we extract the embeddings and optimize the hyperparameters of each individual neural network. Finally we use a weighted averaging method to combine the model contributions based on the performance on a separate holdout set.

3.6 Results

The evaluation dataset consisted of tweets related to personal health experience mentions from influenza and another disease domain, Zika virus. The training data had tweets from two different contexts - user is sick with influenza and user is getting treatment for influenza. The Zika virus has contexts - the user is changing their travel plans in response to Zika concerns, or the user is minimizing potential mosquito exposure due to Zika concerns.

Evaluation criteria: The models are evaluated with respect to Accuracy, F1 score, Precision and Recall. Accuracy is the count of correct predictions. The precision is

Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.88	0.87	0.87	0.86
health concern condition 1	0.92	0.92	0.94	0.90
health concern condition 2	0.75	0.72	0.73	0.70
health concern condition 3	0.91	0.87	0.81	0.94

Table 6: Scores on the Test Set for the Final Model

the ratio tp/(tp + fp) where tp is the number of true positives and fp the number of false positives. The recall is the ratio tp/(tp + fn) where fn is the number of false negatives. The F1 score can be interpreted as a weighted average of the precision and recall[7].

Table 7 describes comparison with other teams. Table 6 describes the scores from this final ensemble model. Our ensemble model surpasses the top performing system from table 7 without using any additional resources.

3.6.1 Analysis

We also ran the BioBERT based models separately on the test set. Table 8 shows the results for using the BioBERT with original embedding. Table 9 shows scores for the BioBERT with generalized embedding. It is interesting to see that the generalized embedding model performs better on the health condition 3 but worse on the health condition 2 thereby proving the effect of using generalizing tokens. The original

[7] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

Team	Accuracy	F1	Precision	Recall
Health concerns in all contexts				
UZH	0.8772	0.8727	0.8392	0.9091
ASU1	0.8456	0.8036	0.9783	0.6818
UChicagoCompLx	0.8456	0.7913	0.9286	0.6894
MIDAS@IIITD	0.8211	0.783	0.8932	0.697
TMRLeiden	0.793	0.7256	0.9398	0.5909
CLaC	0.6386	0.4607	0.7458	0.3333
Health concerns in Context 1: Flu virus (infection/vaccination)				
UZH	0.9438	0.9474	0.9101	0.9878
UChicagoCompLx	0.925	0.9231	0.973	0.878
ASU1	0.925	0.9221	0.9861	0.8659
MIDAS@IIITD	0.8875	0.88	0.9706	0.8049
TMRLeiden	0.8625	0.8493	0.9688	0.7561
CLaC	0.6625	0.5645	0.8333	0.4268
Health concerns in Context 2: Zika virus, travel plans changes				
UZH	0.7536	0.7385	0.7059	0.7742
MIDAS@IIITD	0.6667	0.5818	0.6667	0.5161
ASU1	0.6957	0.5116	0.9167	0.3548
UChicagoCompLx	0.6377	0.4681	0.6875	0.3548
TMRLeiden	0.6377	0.4186	0.75	0.2903
CLaC	0.5362	0.2	0.4444	0.129
Health concerns in Context 3: Zika virus, reducing mosquito exposure				
UZH	0.8393	0.7692	0.75	0.7895
MIDAS@IIITD	0.8214	0.6667	0.9091	0.5263
ASU1	0.8036	0.5926	1	0.4211
UChicagoCompLx	0.8036	0.5926	1	0.4211
TMRLeiden	0.7857	0.5385	1	0.3684
CLaC	0.6964	0.3704	0.625	0.2632

Table 7: System performance for each team submission at the SMM4H 2019 shared task 4

Source: (Weissenbacher et al. 2019)

Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.8947	0.8790	0.9396	0.8257
health concern condition 1	0.9375	0.9367	0.9736	0.9024
health concern condition 2	0.7971	0.7500	0.8400	0.6774
health concern condition 3	0.8928	0.8235	0.9333	0.7368

Table 8: Scores on the Test Set for the BioBERT with Original Embedding Model

Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.8807	0.8650	0.9083	0.8257
health concern condition 1	0.9437	0.9433	0.9740	0.9146
health concern condition 2	0.7101	0.6428	0.7200	0.5806
health concern condition 3	0.9107	0.8648	0.8888	0.8421

Table 9: Scores on the Test Set for the BioBERT with Generalized Embedding model

embedding model does the exact opposite. So the combination of the two ends up evening out the performance across the health concerns 2 and 3.

Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.8631	0.8528	0.8496	0.8560
health concern condition 1	0.9312	0.9316	0.9493	0.9146
health concern condition 2	0.6521	0.6129	0.6129	0.6129
health concern condition 3	0.9285	0.9047	0.8260	1.0000

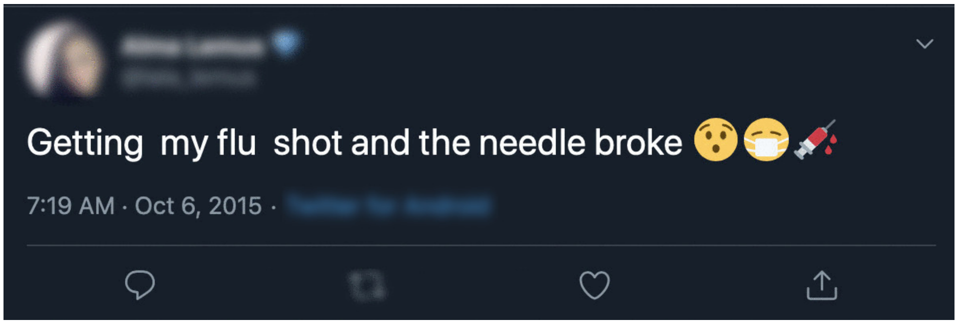
Table 10: Scores on the Test Set using RoBERTa

We also ran the model using RoBERTa representations to see if it will improve the performance. Table 10 shows the results of using RoBERTa in our system. The model improves on health conditions 1 and 3 but the performance is very low on the health condition 2 thereby lowering the overall scores.

According to a 2016 emoji report [8] around 92 percent of the world population use emojis in an online conversation. As such emojis present a very creative and powerful way to communicate so much more than we could with words alone. Use of Emojis has also proven to increase engagement with the rest of the online population. As online communication is increasingly making use of emojis we decided to run another model with that use the information conveyed by the emojis. Instead of removing the emojis during the preprocessing step we convert an emoji present in the tweet into its equivalent textual meaning. We use the emoji package in python3 [9] to convert the emoji to text. Table 10 shows the results on the test set using the emoji to text translation.

[8] http://cdn.emogi.com/docs/reports/2016_emoji_report.pdf

[9] https://pypi.org/project/emoji/



[redacted]

Getting my flu shot and the needle broke 😯😷💉

7:19 AM · Oct 6, 2015 · [redacted]

Figure 5: Tweet with emoji

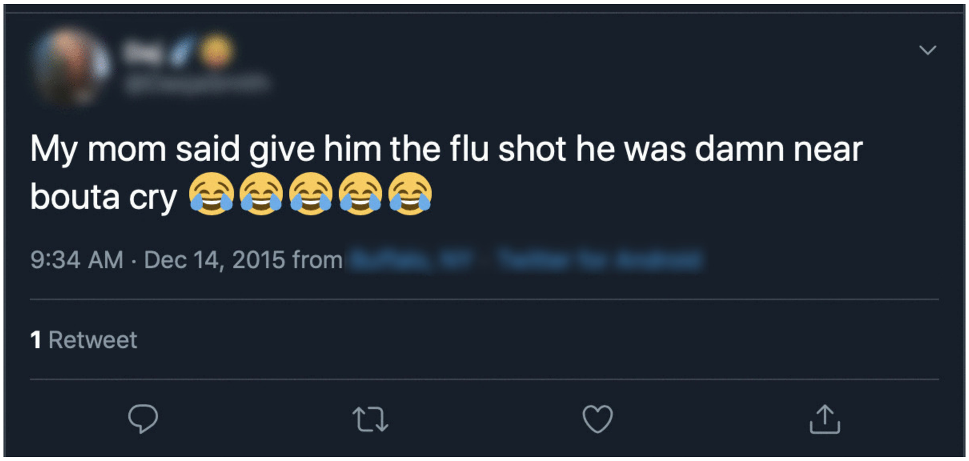
Test set	Accuracy	F1 score	Precision	Recall
health concern overall	0.8631	0.8560	0.8345	0.8787
health concern condition 1	0.9000	0.9036	0.8928	0.9146
health concern condition 2	0.7536	0.7384	0.7058	0.7741
health concern condition 3	0.8928	0.8500	0.8095	0.8947

Table 11: Scores on the Test Set by Emoji Replaced Model

Figure 4. depicts an example of a tweet containing emojis. After converting the emojis to text the tweet becomes

Getting my flu shot and the needle broke :*hushed_face* :: *face_with_medical _mask* :: *syringe*:

Out of the two datasets provided to us - the flu-vaccination and flu-infection tweets, only the former contained tweets with emojis. So we only process the flu-vaccination tweets to replace the emoji with the text. Adding emoji text to the data increases the amount of information available for the model to process but it might not be good enough. The performance of the model is certainly affected when you compare with the model where we remove the emojis.


[redacted]

My mom said give him the flu shot he was damn near bouta cry 😂😂😂😂😂

9:34 AM · Dec 14, 2015 from [redacted]

1 Retweet

Figure 6: Tweet with irrelevant emoji

For example figure 5. depicts a tweet with a lot of laughing emojis. The tweet gets translated as

My mom said give him the flu shot he was damn near bouta cry :***face_with_tears_of_joy***::***face_with_tears_of_joy***::***face_with_tears_of_joy***::***face_with_tears_of_joy***::***face_with_tears_of_joy***:

While analyzing the performance of the model, there were certain tweets that were present in the dataset that were ambiguous. For example in the following tweets the scope of personal mentions is unclear. In the first example the author talks about their daughter having flu and so tweet has been labeled as positive in the dataset. Whereas in the second tweet the author talks about their mother having flu and that is being labeled as negative.

1. *My daughter has developed the flu so I'm home with her. I'm thinking of putting on 'Star Wars' so I can make a joke about R2D2H1N1.*

2. *Is really worried about my mother, she's been diagnosed with the swine flu!"*

There were also some cases which don't have personal mentions but were still labeled as positive like *Watch An Astronaut Give Himself a Flu Shot in Space via http://t.co/feDNZjZATX http://t.co/HGMZ09MmaH http://t.co/vN5CFkXAWJ*

Such cases would not have been an issue if the dataset had a lot more tweets as deep learning models require a lot of data to train. Another case where the model fails is when the tweet is too short. Language models like BERT use bidirectional context to represent the sentence. So in case lack of context affects the performance.

For example:

Just got the flu shot, *23 flu shots today.*, *Time to get a flu shot*

Chapter 4

AUTOMATIC CLASSIFICATIONS OF ADVERSE EFFECTS MENTIONS IN TWEETS

An **adverse drug reaction** (ADR) is an injury caused by taking medication. Adverse drug event (ADE) refers to any injury occurring at the time a drug is used, whether or not it is identified as a cause of the injury. An ADR is a type of ADE whose cause can be directly attributed to a drug and its physiologic properties (Schatz and Weber 2015). A main distinction between ADRs and ADEs is that ADRs occur despite appropriate prescribing and dosing, whereas ADEs may also be associated with inappropriate use of the drug or other confounders that occur during drug therapy but are not necessarily caused by the pharmacology of the drug itself.

4.1 Dataset Description

The data set provided by the SMM4H organizers consisted of 25,672 tweets. Only 2,374 from the entire dataset belonged to positive class so there is a massive class imbalance. For each tweet, the publicly available data set contains: (i) the user ID, (ii) the tweet ID, and (iii) the binary annotation indicating the presence or absence of ADRs.

The tweets were downloaded using the script provided by the organizers of the shared task. Out of 25,672 only 16,220 were available. The class distribution of the downloaded tweets is 1407 positive class tweets and 14813 negative class tweets. The test data had a total of 4575 tweets.

Example:

Positive: @ProfTimNoakes Tim pre diabetic and insulin resistant. My endocologist has prescribed Victoza injections.My **appetite seems to be suppressed**.

Negative: Anyone any experience with Eylea for DME any better than Lucentis. Been offered both just wondering which is best

4.2 Preprocessing

The preprocessing library Ekphrasis provides several methods to clean and process social media text (Baziotis, Pelekis, and Doulkeridis 2017). This library is used to preprocess all the tweets.

- All @user mentions were replaced by user token.
- All HTTP URLs were replaced by url token.
- Hashtags were preprocessed by removing the # symbol and keeping the words.
- Emojis, dates, numbers, etc. are removed.
- The text is converted to lowercase.

4.3 Traditional Machine Learning Approach

We used n-grams as features for our baseline model. We experimented with unigrams and bigrams, and also with the number of features. The graph below represents the scores vs the number of features used in the logistic regression model.

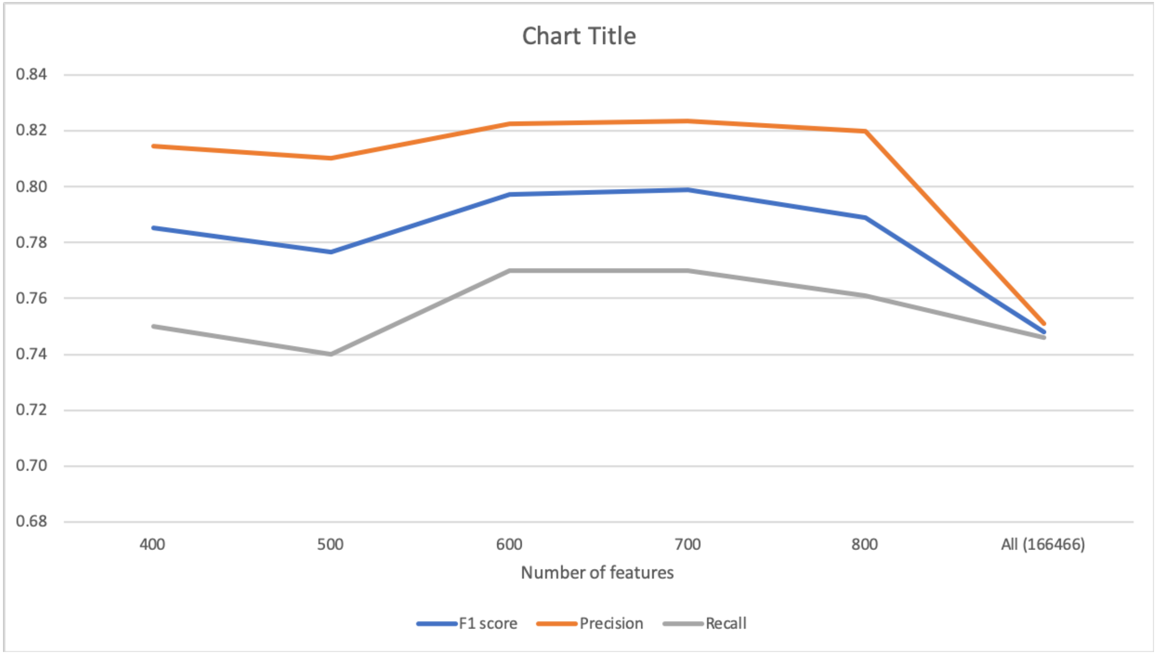


Figure 7: Number of Features Vs Accuracy, Recall, F1 Score

4.4 Transfer Learning Approach

We tried both the fine-tuning based and feature embeddings approaches. We fine-tuned the models on the training dataset and ran two end-to-end approaches. Next we extracted feature representations and used that as input to a separate neural network model. We experimented with the following models

- BERT fine-tune
- BioBERT fine-tune
- BioBERT + FF: CLS token embedding from last 4 layers concatenated + feed forward + softmax
- BioBERT + FF:CLS token embedding from last layer + feed forward + softmax

4.4.1 Data Augmentation

In the training set only 8 percent of total tweets that are ADR related, which indicates a massive class imbalance.So to increase the performance of the model and to make the system more robust we use external datasets that document adverse drug reactions online. We use the CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al. 2015) and Psychiatric Treatment Adverse Reactions (psyTAR) (Zolnoori et al. 2019) datasets. CADEC dataset consists of total 1250 text samples and each sample consists of few sentences. The authors have extracted four entities - ADR, drug, symptom, disease. We use the ADR entity and extract the sentence corresponding to the ADR term. We extracted a total of 2407 such instances.

psyTAR dataset consists of drug reviews posted by people on an online forum. The reviews are split into individual sentences and each sentence is annotated for Adverse drug reactions(ADRs), Withdrawal Symptoms (WDs), Sign/Symptoms/Illness (SSIs) and Drug Indications (DIs). We use the ADR annotations to create additional training samples. We only use sentences which consist of at least 10 words. For example there are sentences like *"Loss of memory", "Very bad side effects.","emotionless, Zombie effect."*. Adding such phrases may make the additional dataset inconsistent with the training set. Some of the sentences do not have ADR mentions we still include them as negative examples.

4.4.2 Pre-training

Pre-training of the language model requires a large corpora, while pre-training is not always necessary but the language construct of tweets differs from that of the

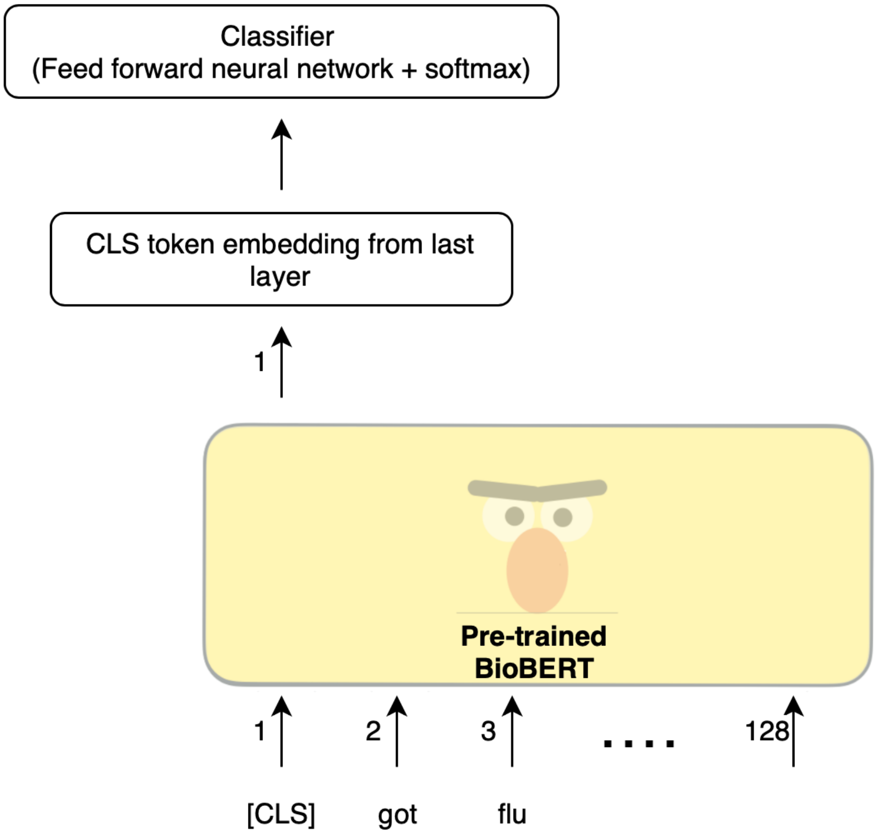

Classifier
(Feed forward neural network + softmax)
CLS token embedding from last layer
1
Pre-trained BioBERT
1
2
3
....
128
[CLS]
got
flu

Figure 8: ADR Detection Model Architecture

Source: Figure adapted from (Alammar 2018)

biomedical text or the Wikipedia corpus. So to collect data for pre-training we use the Twitter API to search tweets by using list of drugs[10] as search keyword. We collected around 0.75 million tweets. We then preprocess the tweets by removing the user mentions and HTTP URLs along with the RT mentions. We pre-train the BioBERT model we set the next sentence prediction loss to 0 so just the masked language modeling task contributes to the overall loss required for the model to learn.

[10] The list of drugs used is compiled by (Sarker and Gonzalez 2015)

4.5 Results

The evaluation metric was micro-averaged F1-score for the ADR class. The logistic regression model was submitted to the SMM4H challenge. Table 8 shows the comparison between various systems submitted to the SMM4H 2019 shared task 1. Table 9 shows the results on the test set with our models. All of our submissions are after the official competition deadline. We did not submit any BioBERT based model during the official competition time frame and all our models have been evaluated in the post-evaluation period. Though our F1 and recall scores are less it is worth noting that our models have been trained on a smaller dataset. Out of 25K tweets we were able to download only about 16K tweets while some of the teams had access to almost all of the tweets.

4.5.1 Analysis

Since the dataset is highly imbalanced and has very less samples of positive class, the performance of the model can be improved with additional examples of positive data as seen in the results.

During analysis of the final model's performance on specific tweet we found that some tweets which just have medical terms get classified as positive. For example:

1. *Disponible SAXAGLIPTINA 5 MG CAJX14 COMP ONGLYZA*

2. *@Senor_Andrew, if you can tell me what Augmentin, Vibramycin, Actos, Fosamax, Paxil and Lipitor... are you have my blessing. #TAMUKNation*

Team	F1	Precision	Recall
ICRC	**0.6457**	0.6079	0.6885
UZH	0.6048	0.6478	0.5671
MIDAS@IIITD	0.5988	0.6647	0.5447
KFU NLP	0.5738	**0.6914**	0.4904
CLaC	0.5738	0.5427	0.6086
THU NGN	0.5718	0.4667	**0.738**
BigODM	0.5514	0.4762	0.655
UMich-NLP4Health	0.5369	0.5654	0.5112
TMRLeiden	0.5327	0.6419	0.4553
CIC-NLP	0.5209	0.6203	0.4489
UChicagoCompLx	0.4993	0.4574	0.5495
SINAI	0.4969	0.5517	0.4521
nlp-uned	0.4723	0.5244	0.4297
ASU BioNLP	0.4317	0.3223	0.6534
Klick Health	0.4099	0.5824	0.3163
GMU	0.3587	0.4526	0.2971

Table 12: System performance for each team submission for the SMM4H 2019 shared task 1

Source: (Weissenbacher et al. 2019)

As we have seen in the previous chapter the performance of the model is poor on short tweets, we find the issue is also present here. For example the tweet *Thank God for trazodone.* predicted as positive.

We also experimented with RoBERTa and table 13 shows that the performance of the model slightly lower than BioBERT. We did expect BioBERT to perform better than RoBERTa because the data had a lot of biomedical terms and BioBERT being trained on a lot of biomedical data.

System	**Data Augmentation**	**F1**	**Precision**	**Recall**
Logistic Regression (Baseline)	-	0.4317	0.3223	0.6534
Average system scores	-	0.5019	0.5351	0.5054
BERT	-	0.5633	0.5173	0.6182
BioBERT	-	0.5814	0.5280	0.6469
BioBERT(last 4) + FF	-	0.5905	0.5420	0.6485
BioBERT(last) + FF	-	0.5912	0.5478	0.6421
Ensemble BioBERT + FF	Yes	0.6170	0.5832	0.6549
Ensemble BioBERT Pretrained + FF	Yes	**0.6430**	**0.6214**	**0.6661**
Ensemble RoBERTa + FF	Yes	0.6114	0.5834	0.6421

Table 13: Final ADR Classification System Scores

Chapter 5

EXTRACTION OF ADVERSE DRUG REACTIONS IN TWEETS

5.1 Dataset Description

The data set provided by the SMM4H organizers consisted of 2,367 tweets out of which 1,212 positive and 1,155 negative. Some of the tweets have multiple ADR mentions which have been documented as separate ADR extraction examples. The total number of such ADR extractions is 2971.

Data Format:

- Tweet Id: 328059687327109120
- Start: 18
- End: 30
- Type: ADR
- Extraction: light' sleep
- Drug: venlafaxine
- Tweet: another night of 'light' sleep. feel low. subdued. sad. #venlafaxine #day12
- Meddra Code: 10062519
- Meddra Term: poor sleep quality

5.2 Preprocessing

The tweets are cleaned by replacing all the @user mentions by *user* token and HTTP URLs with *url* token. In order to run the BioBERT Named Entity Recognition model the data has to converted into a token, tag format. We use the BERT's basic tokenizer that removes white spaces and punctuation to tokenize the tweet. We use the provided ADR extractions to mark the tokens with B or I-tag. In case of the multi word expressions the B-tag is followed I-tags. All other tokens are marked by O-tag.

For example:

```
high off tramadol and venlafaxine in school omg help im falling asleep here
O    O   O        O   O           O  O      O   O    O  B       I      O
```

5.3 Models

BioBERT NER

BioBERT computes the token level B,I,O probabilities from a single output layer using the representations from the last layer. This is the baseline NER model configuration used by BioBERT.

BioBERT NER CRF

Traditionally BiLSTM-CRF models have been found successful at biomedical NER tasks. Since BioBERT is already bidirectional, we add a CRF layer on top of the representation from the last layer (Miftahutdinov, Alimova, and Tutubalina 2019)

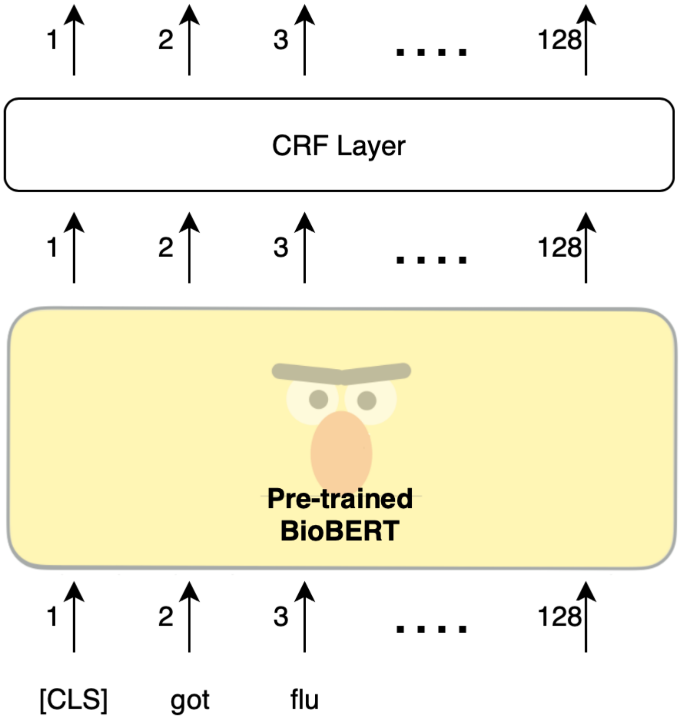


Figure 9: ADR Extraction Model Architecture

Source: Figure adapted from (Alammar 2018)

5.4 Data Augmentation

We use the same two datasets that we had created from CADEC and psyTAR as described in the previous chapter. While CADEC dataset has the ADR spans already extracted with the start and end of the spans, we directly use those to create new samples. psyTAR on the other hand has the extractions specified without the span. So we search the ADR span in the text and extract the start and end indices. If there are multiple ADRs extracted, each ADR becomes part of a separate example in our dataset, this is done in accordance with the multiple ADR case in the training dataset.

We generated 528 such samples from the CADEC dataset and 438 from the psyTAR ADR dataset.

5.5 Pre-trained Model

To improve our scores in the task we use our pre-trained model from the previous task to classify tweets regarding adverse effect mentions. We use the same model BioBERT NER CRF with these pre-trained weights.

5.5.1 Results

The evaluation criteria uses two modes - strict and overlapping. Under strict mode of evaluation, ADR spans are considered correct only if both start and end indices match with the indices in the gold standard annotations. Under overlapping mode of evaluation, ADR spans are considered correct only if spans in predicted annotations overlap with the gold standard annotations. Table 10 describes the system scores for the submissions to the SMM4H shared task 2. Table 8 describes the performances of our models. We did not submit any model during the official competition time frame and all our models have been evaluated in the post-evaluation period.

5.5.2 Analysis

During analysis we found a couple of interesting cases that were present in the dataset.

Multiple ADR Mentions

The following tweet contains more than one ADR mention. The dataset contains such tweets with multiple ADRs split into individual samples.

*"@misterak47 @cbs i don't know what that has to do w/ me. avelox has hurt my **connective tissue,lungs and thyroid**. i guess i should feel lucky"*

For example there are three different ADRs - connective tissue,lungs and thyroid, so there are three individual cases in the dataset.

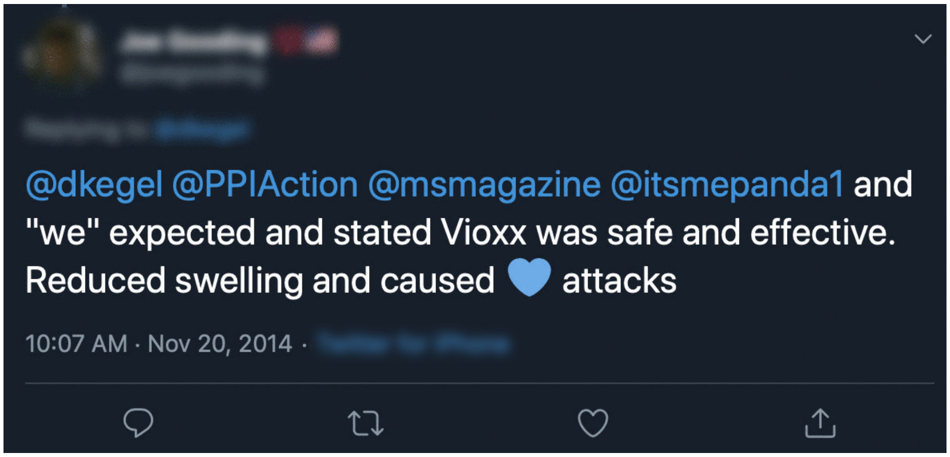
347950374063329280 77 94 ADR **connective tissue** avelox "@misterak47 @cbs i don't know what that has to do w/ me. avelox has hurt my **connective tissue**,lungs and thyroid. i guess i should feel lucky" 10061087 connective tissue disorder

347950374063329280 95 100 ADR **lungs** avelox "@misterak47 @cbs i don't know what that has to do w/ me. avelox has hurt my connective tissue,**lungs** and thyroid. i guess i should feel lucky" 10025082 lung disorder

In this example it is our assumption that dataset creators have divided the ADRs based on the comma and conjunctions like 'and', 'or'. When we pass this example as input to our model it returns connective tissue,lungs and thyroid as one single ADR. We found there were many such examples in the dataset. So we do an additional post processing to handle such cases and split the ADR extraction into multiple spans if possible.

ADR Mention in Emojis

In case of the above tweet the ADR is heart attack. It is spelled by using heart emoji for the word heart and attacks is written normally. Now such cases are not handled by the system, we get partial ADR detection for attacks, thereby affecting our strict F1 scores.



[redacted]

[redacted]

@dkegel @PPIAction @msmagazine @itsmepanda1 and "we" expected and stated Vioxx was safe and effective. Reduced swelling and caused 💙 attacks

10:07 AM · Nov 20, 2014 · [redacted]

Figure 10: ADR Mention with Emoji

Team	Relaxed			Strict		
	F1	Precision	Recall	F1	Precision	Recall
KFU NLP	**0.658**	0.554	**0.81**	**0.464**	0.389	**0.576**
THU NGN	0.653	0.614	0.697	0.356	0.328	0.388
MIDAS@IIITD	0.641	0.537	0.793	0.328	0.274	0.409
TMRLeiden	0.625	0.555	0.715	0.431	0.381	0.495
ICRC	0.614	0.538	0.716	0.407	0.357	0.474
GMU	0.597	0.596	0.599	0.407	0.406	0.407
SINAI	0.574	**0.632**	0.527	0.336	0.37	0.307
HealthNLP	0.542	0.612	0.486	0.36	**0.408**	0.322
ASU BioNLP	0.535	0.415	0.753	0.269	0.206	0.39
Klick Health	0.396	0.416	0.378	0.194	0.206	0.184

Table 14: System performance for each team submission at the SMM4H 2019 shared task 2

Source: (Weissenbacher et al. 2019)

Model	Relaxed			Strict		
	F1	Precision	Recall	F1	Precision	Recall
BioBERT	0.632	0.745	0.549	0.447	0.531	0.385
BioBERT + CRF	0.640	0.733	0.568	0.453	0.523	.400
Pre-trained BioBERT + CRF	0.657	**0.731**	0.596	**0.470**	**0.528**	0.424
RoBERTa	0.617	0.727	0.535	0.430	0.511	0.372

Table 15: ADR Extraction System Scores

Chapter 6

CONCLUSION AND FUTURE WORK

Chapters 3,4 and 5 have described the systems developed to tackle various tasks. Section 6.1 summarizes the work that is done and the contributions made. Section 6.2 discusses some possible future directions.

6.1 Summary

Recent progress in Natural language processing has been due to the rise in transformer models like BERT and other BERT-based models.

The system we used for chapter 3 shows that language models like BERT and BioBERT can be fine-tuned on a small dataset of tweets and still achieve promising results where the health concern is different from the training set.

The results obtained in chapter 4 and 5 show that language models are not enough on their own. It was interesting given that these models are trained on Wikipedia and biomedical text that how well they perform on tweets as tweets often contain misspellings, sarcasm, and slang. So we pre-trained the models on our collection of tweets which improved the performance slightly on the ADR detection task but not so much on the ADR extraction. ADR detection and extraction on electronic health records have much better performance compared to social media data and it still remains a challenge and an opportunity for future research.

The thesis makes the following contributions:

1. Use of term generalization to handle domain adaptation which has not been done any official submission to the generalizable systems task.

2. Use of data augmentation along with pre-training to improve the performance of ADR detection and extraction tasks.

3. Conducted experiments on different models and datasets to compare their abilities.

4. Achieved state of the art accuracy in the generalizable systems task.

5. Developed various models using BioBERT representations which can be reused for similar tasks.

6.2 Future work

BERT and BERT based models require a huge amount of computing power and a large corpus to pre-train the model. To pre-train a model on tweets, we require a much larger dataset than the one used in chapters 4 and 5. Recently compressed models like DistilBERT (Sanh et al. 2019) have managed to achieve similar performances like BERT on various tasks. DistilBERT is 40 percent smaller and 60 percent faster than BERT. Pre-training DistilBERT on a larger corpus of tweets may help us to achieve similar or even better performance on the tasks while managing to minimize the compute required.

Another way to improve the models would be to try and combine the language model representations and other syntactic, lexical, semantic features. Since ADR data on social media is highly imbalanced it would be interesting to see if GPT-2 model (Radford et al. 2019) trained on tweets would be able to generate some positive examples of the data which can be used to further improve the system. Additionally using external knowledge especially in the ADR detection and ADR extraction tasks by combining knowledge from various ADR databases that exist along with language models we may get a better performing system.

REFERENCES

Alammar, Jay. 2018. "The Illustrated BERT, ELMo, and co." *How NLP Cracked Transfer Learning).*

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis. 2017. "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* 747–754. Vancouver, Canada: Association for Computational Linguistics, August.

Bengio, Yoshua, Aaron Courville, and Ian J Goodfellow. 2016. "Deep learning: adaptive computation and machine learning." *Bengio. A. Courville.*

De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. "Predicting depression via social media." In *Seventh international AAAI conference on weblogs and social media.*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805.*

Ge, Suyu, Tao Qi, Chuhan Wu, and Yongfeng Huang. 2019. "Detecting and Extracting of Adverse Drug Reaction Mentioning Tweets with Multi-Head Self Attention." In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task,* 96–98.

Godin, Fréderic, and Baptist Vandersmissen. 2015. "Wesley De Neve, and Rik Van de Walle. 2015." *Multimedia lab acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations. ACL-IJCNLP:* 146–153.

Gupta, Shashank, Manish Gupta, Vasudeva Varma, Sachin Pawar, Nitin Ramrakhiyani, and Girish Keshav Palshikar. 2018. "Multi-task learning for extraction of adverse drug reaction mentions from tweets." In *European Conference on Information Retrieval,* 59–71. Springer.

Hakala, Kai, Farrokh Mehryary, Hans Moen, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2017. "Ensemble of Convolutional Neural Networks for Medicine Intake Recognition in Twitter." In *SMM4H@ AMIA,* 59–63.

Huang, Xiaolei, and Michael J Paul. 2018. "Examining Temporality in Document Classification." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 694–699.

Karimi, Sarvnaz, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. "Cadec: A corpus of adverse drug event annotations." *Journal of Biomedical Informatics* 55:73–81. doi:10.1016/j.jbi.2015.03.010.

Karisani, Payam, and Eugene Agichtein. 2018. "Did you really just have a heart attack?: towards robust detection of personal health mentions in social media." In *Proceedings of the 2018 World Wide Web Conference on World Wide Web,* 137–146. International World Wide Web Conferences Steering Committee.

Kiritchenko, Svetlana, Saif M Mohammad, Jason Morin, and Berry de Bruijn. 2018. "NRC-Canada at SMM4H shared task: classifying Tweets mentioning adverse drug reactions and medication intake." *arXiv preprint arXiv:1805.04558.*

Kiritchenko, Svetlana, Saif Mohammad, Jason Morin, and Berry de Bruijn. 2017. "NRC-Canada at SMM4H Shared Task: Classifying Tweets Mentioning Adverse Drug Reactions and Medication Intake." In *SMM4H@AMIA.*

Kuhn, Michael, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2015. "The SIDER database of drugs and side effects." *Nucleic acids research* 44 (D1): D1075–D1079.

Lamb, Alex, Michael J Paul, and Mark Dredze. 2013. "Separating fact from fear: Tracking flu infections on twitter." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 789–795.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. "Biobert: pre-trained biomedical language representation model for biomedical text mining." *arXiv preprint arXiv:1901.08746.*

Lee, Kathy, Ankit Agrawal, and Alok Choudhary. 2013. "Real-time disease surveillance using twitter data: demonstration on flu and cancer." In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,* 1474–1477. ACM.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. "Opinion observer: analyzing and comparing opinions on the web." In *Proceedings of the 14th international conference on World Wide Web,* 342–351. ACM.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “Roberta: A robustly optimized bert pretraining approach.” *arXiv preprint arXiv:1907.11692.*

Miftahutdinov, Zulfat, Ilseyar Alimova, and Elena Tutubalina. 2019. “KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue.” In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task,* 52–57.

Neiger, Brad L, Rosemary Thackeray, Scott H Burton, Callie R Thackeray, and Jennifer H Reese. 2013. “Use of twitter among local health departments: an analysis of information sharing, engagement, and action.” *Journal of medical Internet research* 15 (8): e177.

Nikfarjam, Azadeh, Abeed Sarker, Karen O’Connor, Rachel Ginn, and Graciela Gonzalez. 2015. “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features.” *Journal of the American Medical Informatics Association* 22 (3): 671–681.

Paul, Michael J, and Mark Dredze. 2011. “You are what you tweet: Analyzing twitter for public health.” In *Fifth International AAAI Conference on Weblogs and Social Media.*

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “Glove: Global vectors for word representation.” In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* 1532–1543.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep contextualized word representations.” *arXiv preprint arXiv:1802.05365.*

Prieto, Vıctor M, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luıs Oliveira. 2014. “Twitter: a good place to detect health conditions.” *PloS one* 9 (1): e86191.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners.”

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” *arXiv preprint arXiv:1910.01108.*

Sarker, Abeed, and Graciela Gonzalez. 2015. “Portable automatic text classification for adverse drug reaction detection via multi-corpus training.” *Journal of biomedical informatics* 53:196–207.

Schatz, S, and R Weber. 2015. “Adverse drug reactions.” *Pharmacy Practice* 1:1.

Shen, Chen, Hongfei Lin, Kai Guo, Kan Xu, Zhihao Yang, and Jian Wang. 2018. “Detecting adverse drug reactions from social media based on multi-channel convolutional neural networks.” *Neural Computing and Applications:* 1–10.

Speer, Robert, Joshua Chin, and Catherine Havasi. 2017. “Conceptnet 5.5: An open multilingual graph of general knowledge.” In *Thirty-First AAAI Conference on Artificial Intelligence.*

Wang, Alex, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. “Glue: A multi-task benchmark and analysis platform for natural language understanding.” *arXiv preprint arXiv:1804.07461.*

Weissenbacher, Davy, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O’Connor, Michael Paul, and Graciela Gonzalez. 2019. “Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019.” In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task,* 21–30.

Xherija, Orest. 2018. “Classification of medication-related tweets using stacked bidirectional LSTMs with context-aware attention.” In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task,* 38–42.

Yin, Zhijun, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. 2015. “A scalable framework to detect personal health mentions on Twitter.” *Journal of medical Internet research* 17 (6): e138.

Zolnoori, Maryam, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D Shah, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, et al. 2019. “The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications.” *Data in brief* 24:103838.

APPENDIX A

ABLATION STUDIES

A.1 Fine-tuning parameters

Fine-tuning on smaller datasets results in a variance in the evaluation accuracy. The parameters responsible for this variance are the learning rate, number of epochs, and batch size.

For the BioBERT based models in chapter 2 and 3, we ran an exhaustive search on these parameters with the following range criteria:
Learning rate: 5e-5, 3e-5, 2e-5
Batch size: 8,16,32
Number of epochs: 2,3,4,5

For chapter 4 since the model runs end-to-end the only parameters that we varied was the number of epochs.

Number of Epochs	Relaxed F1 score	Strict F1 score
8	0.677	0.522
10	0.667	0.534
16	0.637	0.519
20	0.662	0.473

Table 16: Number of epochs vs F1 scores

A.2 Embeddings

We experimented with different combination of masked embeddings
1. Average embedding of all tokens
2. Embedding of the CLS token from each of the last two individual layers
3. Average of CLS token embeddings from last 4 layers

A.3 Pre-training steps

This ablation study answers the following two questions:

Masked Embedding	F1 score
Average of all token embeddings	0.88
CLS token from last layer	0.902
CLS token from second-to-last layer	0.893
Average of CLS token from last 4 layers	0.881

Table 17: Embedding configuration vs F1 scores

1. Will evaluation accuracy increase after k training steps?

2. Will the masked language modeling task work effectively on a smaller sequence and less number of masked tokens?

Figure 8 shows the masked language accuracy (on dev set) which increases and the loss decreases as the number of training steps increase.

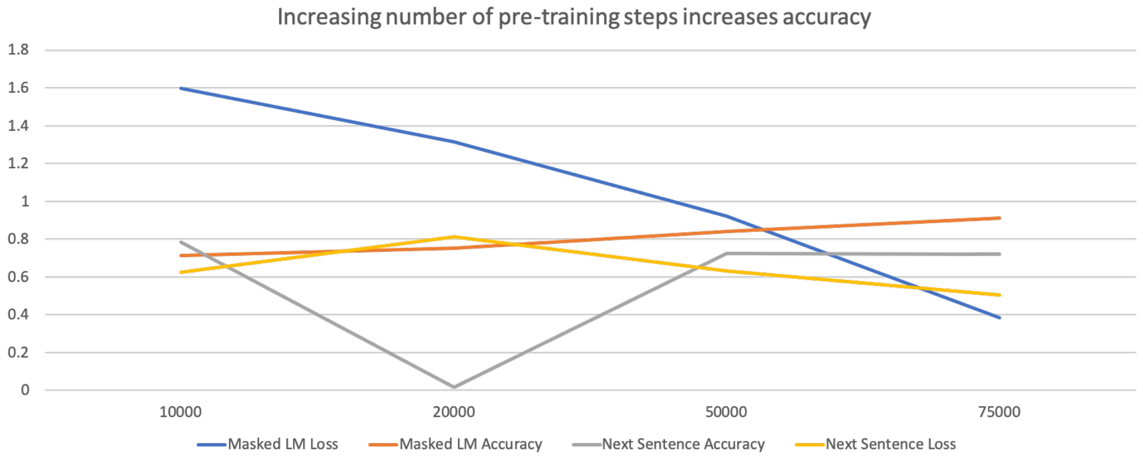


Figure 11: Loss and accuracy for pre-training tasks

APPENDIX B

LIST OF TERMS USED FOR MASKING FLU-MENTIONS

- Swine
- flue
- colds
- tonsillitis
- cold/flu
- hayfever
- tonsilitis
- flu
- h1n1
- h5n1
- influenza
- viruses
- Virus
- fluey
- flulike
- flunami
- asian flu
- stomach flu
- superflu
- bird flu
- avian flu
- tonsilitis
- strep
- sinusitis
- bronchitis

APPENDIX C

NLP PIPELINE IMPLEMENTATION

Fine-tuning parameters:

```
python biobert/run_classifier.py \
  --task_name=COLA \
  --do_train=true \
  --do_eval=true \
  --data_dir=data \
  --vocab_file=biobert_v1.1_pubmed/vocab.txt \
  --bert_config_file=biobert_v1.1_pubmed/bert_config.json \
  --init_checkpoint=/../model.ckpt \
  --max_seq_length=128 \
  --train_batch_size=32 \
  --learning_rate=5e-5 \
  --num_train_epochs=5.0 \
  --output_dir=/../task1_output/
```

Pre-training parameters:

```
python biobert/run_pretraining.py \
  --input_file=/../tf_examples.tfrecord \
  --output_dir=/../pretrained2 \
  --do_train=True \
  --do_eval=True \
  --bert_config_file=/../biobert_v1.1_pubmed/bert_config.json \
  --init_checkpoint=/../biobert_v1.1_pubmed/model.ckpt-1000000 \
  --train_batch_size=256 \
  --max_seq_length=64 \
  --max_predictions_per_seq=5 \
  --num_train_steps=100000 \
  --num_warmup_steps=10000 \
  --learning_rate=2e-5
```

The fine-tuning and training our custom BioBERT models was performed on 2 Tesla V100-SXM2-16GB GPUs.

The pre-training of BioBERT was done on 4 Tesla V100-SXM2-16GB for 10 days. We followed the author's instruction regarding the learning rate, training batch sizes. We decided to keep the maximum sequence length to 64 that since tweets are usually of shorter length. The default value of maximum predictions per sequence is 20, and it is likely that a tweet may end up being shorter than 20 tokens and cause an issue with the training process, we decided to reduce that number to 5.

APPENDIX D

CODE REPOSITORY

The Code and intermediate data is available at
https://github.com/ShubhamGondane/Biomedical-NLP-Twitter

APPENDIX E

LIST OF DRUGS

The following list has been compiled by (Sarker and Gonzalez 2015) and was used to collect additional tweets for pre-training.

- humira
- dronedarone
- lamictal
- pradaxa
- paxil
- zoledronic acid
- trazodone
- enbrel
- cymbalta
- quetiapine
- cipro
- lozenge
- dabigatran
- olanzapine
- fluoxetine
- vyvanse
- seroquel
- fosamax
- paroxetine
- nicotine
- effexor
- prozac
- tysabri
- rivaroxaban
- baclofen
- lamotrigine
- venlafaxine
- apixaban
- avelox
- levaquin
- zyprexa
- duloxetine
- ofloxacin
- geodon
- victoza
- metoprolol
- viibryd
- pristiq
- nesina

- factive
- gamma-aminobutyric acid
- sabril
- livalo
- denosumab
- bystolic
- xarelto
- floxin
- boniva
- saphris
- ziprasidone
- memantine
- namenda
- latuda
- fycompa
- canagliflozin
- zometa
- etanercept
- lurasidone
- alendronate
- linagliptin
- effient
- vimpat
- eliquis
- liraglutide
- pregabalin
- onglyza
- nicotrol inhaler
- lyrica
- invokana
- commitlozenge
- actonel
- nicotrolinhaler
- synthroid
- albuterol
- nasonex
- spiriva
- suboxone
- nexium
- januvia
- valsartan
- tamiflu