ORIGINAL RESEARCH PAPER

Engineering

KEY WORDS: COVID-19, Sentiment Analysis, BERT, NLP, Deep Neural Net, Twitter, PyTorch, Data Mining. Opinion Mining

SENTIMENT ANALYSIS ON COVID-19 RELATED TWEETS

Jinesh Patel M.S in Engineering Online University of California Riverside, CA,, USA

1. Introduction

COVID-19[1] is a pandemic going on during the current time. COVID-19 poses a new challenge to humankind. Meanwhile, It is essential to know what the majority of people thinking about this pandemic. Sentiment analysis[2] is a widespread technique applied in field of marketing, customer feedback and consumer research. This project performs sentiment analysis on COVID-19 related tweets[3] and getting some higher idea about people's sentiment toward pandemic. After performing sentimental analysis on COVID-19 related tweets, this project also draws some interesting findings and conclusions in the last section. Performing sentiment analysis on social media data like tweets, Facebook comments are challenging task for some reasons .1) Due to high volume, data may or may not be in csv, excel ,or other text-rich formats. Performing operations on data needs 2) To draw human sentiment from raw text itself complex task for a machine. However, recently there are advanced deep learning based methods available for performing such a complex task. For example, Answering the question by chatbot or auto-compilation of a statement in email.[4]. This project takes NLP(Natural Language Processing) approach with the help of Google's pre-trained deep neural network called BERT[5] for identifying sentiment from raw text data. The layers of deep neural network implemented in PyTorch[6].

Deep learning is an advanced method of machine learning and data mining. It is used in field of image classification[7], In finance like a market prediction or portfolio management. However, in sentimental analysis cases, deep learning becomes challenging because of the variety of human languages. For example, model trained in the English language for finding text sentiment, will not predict sentiment for the Chinese language. As this limitation, In this project, sentiment analysis performed on English language.

Training a deep learning network for predicting sentiment analysis from a raw text is another challenge for the project. Deep learning networks are prone to overfitting in case of insufficient data[7].

1 Dataset for COVID-19 related tweets openly available at this link . It is open source and maintain weekly.
2. Hydrator is an electron based desktop application for downloading tweets from tweet Ids.
3. Generally sentiments are either positive or negative but here we have classified in three classes.

CS235 Data mining course project. 2020.

Training requires a large amount of data. In some cases , training NLP based deep networks needs multimillions words corpus[8]. Model training from scratch is a time consuming complex task.

Also, this trained model entirely from scratch does not give a guarantee of good performance[8]. The second approach for solve this problem is, to use the power of transfer learning and fine tuning. Transfer learning[9] is a powerful technique for training a deep neural network. In this project, this approach is selected.

A major challenge for fine tuning BERT based networks is training millions of weights and biases and of course, this process is time consuming. In this project, the various architectures of fine-tuned model and hyperparameters evaluated. These all experiments are covered in experiment evaluation section. Also, in the evaluation section, there are some experiment results and data comparisons displayed. The model with higher train accuracy and higher test accuracy indicates generalization. This model can be used for predicting sentiment from unknown raw text like "I love data mining" or "I hate long waits for model training". After selecting the best generalized model trained from tagged data (supervised learning), It is used in drawing sentiments from COVID-19 related tweets. Tweets data is not easy to crawl for several reasons. 1) Identification: need to identify which tweets are COVID-19 related

2) Twitter API an upper limit: twitter API has upper limit per user ,so it cannot allow downloading more than specific data per hour. Luckily some data with tweet ids1 is available. Tweet ids are not solving problems because they do not have full tweet data and metadata, To get full tweet data tool called hydrator2 is used in this project.

In last section of this paper, we will see some interesting conclusions about COVID-19 related tweets after tagging sentiment [negative, neutral, positive]3 and discuss the further scope for researching or extending the project.

2 Related Work

2.1 Deep neural network related work

Sentiment analysis is one of the hot topics among researchers and there are several papers and articles written about it. A deep neural network is a popular data mining technique that solves several classification and regression tasks efficiently. BERT paper was published by Google AI Language team Devlin at el 2018[5]. BERT showed data that was superior to previous models and solved complex tasks like completing sentences and answering questions. BERT also showed effectiveness in transfer learning and fine tuning. Some of the tasks were performed by BERT 1) Masking LM. 2) Next sentence prediction, 3) Answering questions.

2.2 Sentimental analysis related work

There are several papers written on sentimental analysis. The closest paper related to the project was Sentiment Analysis of Twitter Data by Agarwal at el[10]. In this paper, the model was trained from scratch by using conventional NLP techniques. Unigram model achieved around 75% accuracy. However, this model was not a use case of transfer learning or fine tuning.

Another noted work related twitter sentiment was published in 2014, titled Twitter Sentiment Analysis by Sarlan at el.[11]. In this paper authors mentioned various sentiment analysis techniques like NLP, Case-Based-Reasoning(CBR) and methods involving ANN(Artificial Neural Network). The model correctly identified sentiments from extensive amount JSON tweet data.

3 Proposed Method

In this section, I am going to describing a detailed approach taken by this project. I am also going to mention some significant issues and challenges faced during the project.

3.1 Supervised learning approach and challenges

Sentiment analysis is a supervised learning task in data mining field where you can train your model with tagged data. Tweets are not tagged dataset. So it is hard to a trained model based on asupervised learning approach. Also, It is difficult to evaluate since there is no ground truth available. To solve this fundamental issue, I used a hybrid approach in this project. There is two-step process for hybrid solution Step 1: train and fine tune model on balanced4 labeled dataset and get the finest model. Step 2: Use this model to predict the sentiment of the unknown raw text (i.e tweets).This

4 In balanced data sets all classes are equally divided. approach will solve the problem related to data incompleteness. However, it will add some bias towards original data. For example, if a model is trained on IMDB movie review dataset. It has a bias toward vocabulary used in movie jargon words, like "blockbuster", "superhit" etc. This is a limitation of this approach.

3.2 Project Architecture

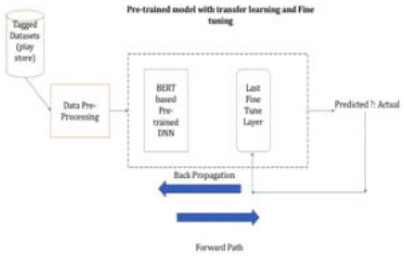

Figure 1: Model architecture with its components and training data flow.

In this section, I am going to explain model implementation in detail and how training works in this project. The next section will discuss the result and some comparisons based on various parameters and hyperparameters. BERT comes with various pre- trained models. These models are implemented by huggingface[12]

.Selection of model highly depends on the use case, For example, the Chinese language based model also available for creating chatbot that understand the Chinese language. In this project, I tried two pre-trained models called bert-based-cased and bert-based- uncased. Both models are pre-trained for English language tasks. This project is implemented in Python programming language with the help of a deep neural network tool called PyTorch that is developed by Facebook AI.

Project data flow starts from tagged datasets. In this project , I tried various datasets for the training model. However, the final selection was google's play store datasets. This dataset has a star ratings with balanced classes. To convert start rating into sentiment, I assumed greater than 4 ratings as a positive and tagged sentiment as "2". 3 ratings as neutral sentiment tagged sentiment as "1". Less than 3 rating as negative sentiment tagged as "0".

BERT is expecting input in a certain format. This is a limitation of BERT. Input data need to be tokenized according to BERT format and this token length is a hyperparameter. By observing training datasets(mostly raw text).This project set maximum token length N=192.BERT also need to insert special tokens like [CLS] for classification, [UNK] for unknown word and [PAD] for padding. Sentiment analysis on COVID-19 related tweets J.Patel, December, 2020, Riverside, California USA

Sentiment analysis on COVID-19 related tweets J.Patel, December, 2020, Riverside, California USAFortunately, this tokenizer also available as part of hugging-face transformers module. These tokens from raw text fed to the BERT input layer and propagate through various hidden layers, and finally pass through fine-tune layers. It gives an output of the probability of sentiment classes. The final output is the maximum of this probability. Since our training data is tagged data, we can measure error through cross entropy. During training, neural networks adjust parameters (weights and biases) via the backpropagation method to minimize this error. Below steps describes the training and fine tuning of BERT based neural net.

1) Tagged training data sets divided into training data set and test data sets by 90%:10% ratio. For validation, the test dataset was further divided into 70%:30% ratio.

2) In training datasets, raw text is converted into tokens and masks. Mask is a special vector that has positional values. E.g. 1 for word present 0 for not.

3) Token propagates through the neural net and passes though the fine tuning layer. Get sentiment probabilities.

4) By comparing actual sentiment, get an estimate of error. By the backward propagation method and find training accuracy and validation accuracy.

5) About E=10 epochs, model shows saturated accuracy on training as well as saturated accuracy on validation set.

This process is repeated with several hypertuning parameters and on couple of fine tune layer arrangements. Experiments results shown in next section. The best accurate model selected for predicting sentiment on tweets.

Model implementation is done in python programming language with help of huggingface transformer library and PyTorch. Some of the experiment reference taken from this[13] book.

4 Experiments and Results

Several experiments were conducted during this project. These experiments were based of mainly three prospects. The model has also selected based on these experiment results. All model experiment were conduct on AWS(Amazon Web Service) based g3.8xlarge instance with NVIDIA Tesla M60 GPU. The training set raw text contains around 15K sentences with avg sentence length of around 150 words per sentence(this makes training dataset total of 2.25 million words).As this problem is a classification problem, In all experiments, cross entropy function[14] is used as loss function.

4.1 Pre-trained model

BERT comes with various pre-trained models; these pre-trained models are trained on various vocabulary. The selection of these pre-train models plays an important role in model accuracy and functionality. For example, BERT comes with a large pre-trained model that covers a greater amount of English language vocabulary. However, training and tuning of such a large model is challenging and time-consuming work. Most of the GPU based machine is run out of memory to train this model. So, as per pre- trained model selection, the experiment was conducted on mainly two pre-trained models 1) bert-base-cased (this model has a case sensitive data) 2) bert-base-uncased. Experiment results are displayed below in figure 2.

Pre-trained model	Training accuracy	Validation accuracy	Test accuracy
bert-base-cased	0.9801	0.8818	0.8807
bert-base-uncased	0.9775	0.8691	0.8821

Figure 2: Experiment result based on pre-trained model selection.

4.2 Fine tuning layer

The fine tuning layer is another interesting aspect for this project, linear used as fine tune layer. Three sets of architectures used in evaluation of model 1) Simple: just linear layer connected with last hidden layer of BERT. 2) With drop out 35%: deep neural network prune to overfitting so dropout layer generally used to avoid this situation. 3) Two layers with dropout: instead of one linear layer, two layers used with dropout 35%. Experiment results are displayed below in figure 3.Here, pre-trained model selected was bert-cased based on higher training/validation accuracy.

Layer	Training accuracy	Validation accuracy	Test accuracy
Simple	0.9786	0.8703	0.8858
With drop out 35%	0.9810	0.8754	0.8952
With drop out 35% 2 layers	0.9801	0.8818	0.8807

Figure 3: Experiment result based on fine tuning layer architecture selection.

4.3 Hyperparameters

Hyperparameters are essential for fine tuning any deep neural network. Selecting correct hyperparameters is important for any machine learning problem. In this project, I have tried various hyperparameters 1) optimizer 2) schedular 3) no of epochs. Experiments show no of epochs is not much effective parameter after N > 10. Training accuracy, as well as validation accuracy, saturated within epochs (see figure 6). This is true for all experiments. The rest of the hyperparameters optimizer and schedular, result shown in figure 4 and figure 5.

Optimizer	Training accuracy	Validation accuracy	Test accuracy
Adam	0.9801	0.8818	0.8807
AdamWeightDecay	0.9757	0.8732	0.8625

Figure 4: Experiment result based on optimizer selection.

Schedular	Training accuracy	Validation accuracy	Test accuracy
get_linear_schedule_with_warmup	0.9801	0.8818	0.8807
get_constant_schedule	0.9784	0.8798	0.8841

Figure 5: Experiment result based on schedular selection.

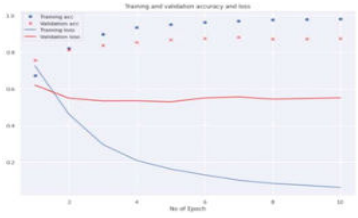

Figure 6:Training/validation accuracy vs No. of epochs.

4.4 Final model.

Based on various experiments conducted, results show most of the models' generalized sentimental analysis task. However, the best model is a combination of bert-cased, 35% dropout layer, Adam optimizer with linear warmup schedular with epochs E=10. Accuracy is not the only metric that should be trusted in the supervised learning task. After selecting the final model, I have performed some other experiments to get confusion matrix, precision, and recall from test data. Figure 7

Figure 7: Confusion matrix of the best selected model.

This confusion matrix shows some classification task on testing data. We can see some misclassified class mostly between neutral sentiment to positive sentiment. By definition lets calculate recall and precision of this model

Recall = True Positive / (True positive + False negative) (1)
Precision = True Positive / (True Positive + False Positive) (2)
F1 score = 2 * (P * R / P + R). (3)

Based on all three formula we are getting model score for Rnegative = 0.85 Rpositive= 0.90 and Rneutral = 0.86 Also, precision we get 0.92, 0.92, 0.80 respectively.

4.5 Twitter sentiment experiment

Fine tune model selection based on pretrained model is process based on various experiments and results. Twitter data related to COVID-19 downloaded and randomly sampled based on time the people tweets the most[15].. This data is obtained between Jan-2020to Oct-2020.100K tweets selected randomly for each month and performed sentiment analysis on raw tweet text data.

Found some interesting fact after performing analysis.

Conclusion 1: Most people have negative sentiment in COVID-19 related tweets. You can see in figure 8, sentiment class distribution based on 100K tweets.

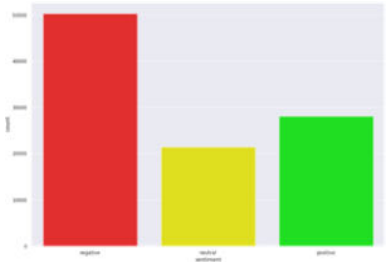

Figure 8: sentiment distribution based on 100K tweets.

Conclusion 2: People who have verified accounts, are less likely to tweet negatively about COVID-19.

This is an interesting finding, after analyzing tagged data with the sentiment. Finding that, on verified accounts, sentiment class distribution is almost 50%-50% for negative and positive classes. Compare to the overall distribution of 50%-30%.

Conclusion 3: People with more number of followers are less likely to tweet negatively about COVID-19.

Conclusion 4: People with negative sentiment tweets are highly biased towards government agencies like CDC.

Conclusion 5: Word cloud is a visualization technique for NLP. It displays most frequent word by size. Figure 9 shows

negative sentiment statement word cloud. We can see some highly negative words which are used frequently. Figure 10 shows word cloud for positive sentiment. Some of the words like "China", "Wuhan", "People", "Trump" are common in both clouds. While, you can see some positive words like "Thanks", great" etc are more frequent in positive sentiment.

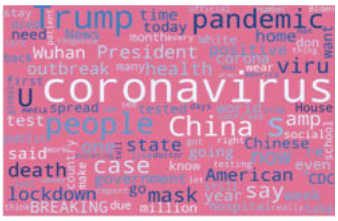

Figure 9: Negative sentiment word cloud.

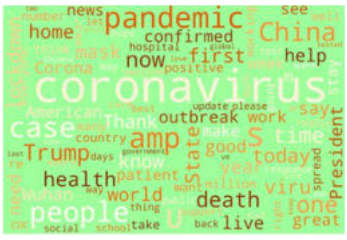

Figure 10: Positive sentiment word cloud.

5 Conclusion

This project is taking a hybrid approach for performing sentiment analysis on COVID-19 related tweets. First, the model is trained via supervised learning on available tagged dataset. This model training is actually part of fine tuning on pre-trained NLP deep neural network called BERT. After selecting the best model, we can use the same model for performing sentiment analysis on the unknown statements.

Initially, plan was to perform sentiment analysis on 3 millions tweets. But due to time consuming process it is limited to 100K tweets. NLP based project is a complex task and needs a greater amount of computing and time. However, once the model trained and tuned. It is used anywhere for performing the same task. In Sentiment analysis on COVID-19 related tweets J.Patel, December, 2020, Riverside, California USA future, this project can be extended to predict sentiment analysis on Facebook comments or product review. Even in twitter, after tagging sentiment of tweets, Data scientist may get useful information from data.

REFERENCES

[1] https://covid.cdc.gov/ COVID-19 related resources and information on CDC website.
[2] https://en.wikipedia.org/wiki/Sentiment_analysis basic definition of sentiment analysis .
[3] This tweets are obtained and marinated by part of USC project github link is here.
[4] https://blog.google/products/search/search-language-understanding-bert/ BERT google blog
[5] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Google AI Language.
[6] PyTorch developed by Facebook AI labs for deep neural net. https://github.com/pytorch/pytorch
[7]. Deep Learning with python Manning publication, François Chollet, November 2017 ISBN 9781617294433.
[8] Agrawal et al. Proceedings of the Workshop on Language in Social Media (LSM 2011), Sentiment Analysis of Twitter Data.
[9] Transfer Learning for Natural Language Processing. Manning publication, Paul Azunre, ISBN 9781617297267.
[10] Agrawal et al. Proceedings of the Workshop on Language in Social Media (LSM 2011), Sentiment Analysis of Twitter Data.
[11] Sarlan at el 2014, Twitter Sentiment analysis.
[12] Hugging face is developing libraries for NLU (natural language understanding) and NLG (natural language generation).BERT transformer library is develop by them https://huggingface.co/transformers/model_doc/bert.html
[13] Get the sht done with pyTorch. Solve Real-World Machine Learning Problems.
[14] Cross entropy loss function definition and explanation https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
[15] STUDYTWEETTIME DISTRIBUTIONINS: https://buffer.com/resources/best-time-to-tweet-research