Sentiment Analysis Task on Twitter Data

Abhishek Akshat

Abstract. In this research paper, we used BERT (Bidirectional Encoder Representations from Transformers). We benchmarked the performance of BERT against current methods to perform the Sentiment Analysis task on the Twitter data.

Keywords: Sentiment Analysis · Twitter · BERT

1 Introduction

The determination of polarity of texts to identify its sentiment has become an important task today in the Natural Language Processing (NLP) and data science field. This is becoming an interesting field to study due to large amounts of social media messages. The tweets from Twitter have especially become an integral part of people's lives who want to express their views through this social media platform in a concise manner within the 140 character limit. Various techniques in NLP are being applied to these tweets to gather more information about people's sentiment. In this research, we are interested in the application of one such technique called Sentiment Analysis.

Pang et al. (2008) [1] defines Sentiment Analysis as follows: "*Sentiment analysis is the process to identify and analyze polarity from short texts, sentences, and documents.*"

According to them, sentiment seems to require a higher level of understanding than just topic-based classification. Sentiment analysis has many uses in the field of political and social sciences as well as for businesses. Companies can use sentiment analysis to analyze customers' satisfaction level of their products and accordingly improve their products based on the opinions of customers to provide better services in the future.

In this research paper, we are going to use BERT (Bidirectional Encoder Representations from Transformers) [2]. We will benchmark BERT against current methods to perform the Sentiment Analysis task on the Twitter data.

2 Related Work

Sentiment analysis has become an interesting field for researchers with the increase in amount of text messages from social media and blog posts. A comprehensive overview of prevailing work has been given in Pang and Lee, 2008 [1]. In their paper, they have described the current approaches and techniques for an

opinion-oriented information retrieval. Pak and Paroubek (2010) [3] in their paper have scraped tweets from Twitter using Twitter API. They combined those tweets together and created a corpus in which each tweet was annotated by emoticons. A Multinomial Naıve Bayes classifier which used N-gram and POS-tags as features was trained and tested on the corpus. Parikh and Movassate (2009) [4] have used Naıve Bayes bigram model to classify tweets and compared it with Maximum Entropy model. They inferred that Naıve Bayes model performs much better than Maximum Entropy model. On the contrary, Go and L.Huang (2009) [5] say that SVM outperforms other models. They made use of unigrams, bigrams and POS for their feature space. Cliche (2017) described in his paper about Twitter sentiment classifier using Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) networks. His sentiment classifier utilized huge volumes of unlabeled data to pre-train word embeddings. A subset of the unlabeled data set was then used to refine the word embeddings using distant supervision. Lastly, the final CNNs and LSTMs were trained on the SemEval-2017 Twitter dataset where the word embeddings are refined again. The goal was to improve performance of his sentiment classifier for which several CNNs and LSTMs were combined together.

3 Data

For this research paper, we have used the SemEval 2017 Tweets dataset [6]. The data collected from Twitter is particularly useful in Sentiment Analysis for the following reasons:

- Micro-blogging platforms such as Twitter are being used by people from a variety of backgrounds to convey their views on diverse topics making tweets a valuable source of public opinion.
- Twitter consists of a large amount of tweets that is growing every passing day making the gathered data set arbitrarily huge.
- The user base of Twitter is diverse ranging from regular users to business executives, celebrities, politicians and country presidents and prime ministers. Hence, it is possible to gather tweets of users from diverse social, political and interest groups.
- Twitter's user base is also characterized by people from different countries

We explore the data in the further sections.

3.1 Data Description

The dataset provided contains 11 *.tsv* files (Tab Separated Files). We merged all the files in one large file in order to perform data pre-processing and tasks easily.

We can observe the sample of our Twitter data in the figure 1. The dataset provided had 3 columns in each file, namely

	Id	Sentiment	Text
0	260097528899452929	neutral	Won the match #getin . Plus\u002c tomorrow is ...
1	263791921753882624	neutral	Some areas of New England could see the first ...
2	264194578381410304	negative	@francesco_con40 2nd worst QB. DEFINITELY Tony...
3	264041328420204544	neutral	#Thailand Washington - US President Barack Oba...
4	263816256640126976	neutral	Did y\u2019all hear what Tony Romo dressed up ...

Fig. 1. Data Sample

- Item ID
- Sentiment (Positive, Neutral, Negative)
- Sentiment Text (Tweet)

4 Method

In this section, we will describe the methods and experiments performed. We used a pre-trained general BERT to perform the task of Sentiment Analysis. To perform the Sentiment Analysis using pre-trained BERT and evaluating and comparing its performance against current methods, we choose **Logistic Regression** and **Support Vector Machines** as our baseline methods.

4.1 Data Pre-processing

The data provided is collected from Twitter hence, it contains some noise and needs to be pre-processed before we can use it to train our models. So we performed several tasks to clean our data.

We removed all the NaN values, after that we removed stop-words from the Tweets. We also converted texts to lowercase and removed punctuation. Text normalization is important for noisy texts, therefore we also performed text normalization for our dataset since Twitter data has noise present.

After cleaning the data, the final dataset contained 48302 tweets. After preprocessing the data, we divided the dataset into training and testing sets. For our experiments, we used 80% of the data for training and the remaining 20% for testing purpose.

The distribution of Tweets with respect to its polarity can be seen in the figure 2. It shows that the polarity is almost fairly distributed among the dataset used for performing the experiments in this paper.

4.2 Experiment

The used BERT model is a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture as we can see from the figure 3. This pre-trained BERT was trained on English Wikipedia (2,500M words) and BooksCorpus (800M words).

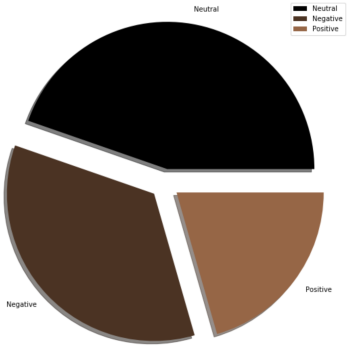

Fig. 2. Sentiment Distribution

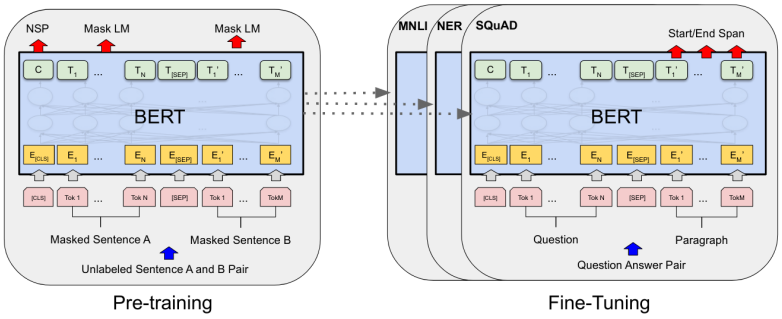

Fig. 3. BERT Model

We choose *'CategoricalCrossEntropy'* as our loss function, *'SparseCategoricalAccuracy'* as our accuracy metric and *'Adam'* as our optimizer for configuring the BERT model. We fine-tuned the model for 5 epochs with the training dataset. Then we evaluated its performance on the testing data.

For baseline comparison, we used two traditional machine learning algorithms namely, Logistic Regression and Support Vector Machines. For our experiments, we trained all our baselines models with the help of some vectorization strategies such as bag-of-words, TF-IDF, Word2Vec, etc. We trained and evalute the baseline models on the same training and testing sets.

We recorded the Precision, Recall and F1-Score for our models. The results from the performed experiments is shown and explained in the next section.

5 Results

After training the models on a collection of 38642 tweets and evaluating it on the test set of 9660 tweets, we recorded the results for our pre-trained BERT, Logistic Regression and Support Vector Machines.

Model	Precision	Recall	F1-Score
Pre-trained BERT	**0.5188**	**0.4961**	**0.5476**
Logistic Regression	**0.4282**	**0.4115**	**0.4062**
Support Vector Machines	**0.4316**	**0.3972**	**0.3880**

From the results table, we can see that pre-trained BERT performs better than the respective baseline models over all metrics. However, Logistic Regression seems to perform better than the Support Vector Machines. SVM seems to perform worse when we have a large dataset and the dataset contains a lot of noise since Twitter data has a lot of noise this can be the reason behind its performance.

6 Conclusion

We successfully built a transformers network with a pre-trained BERT model and achieved good results on the sentiment analysis of the Twitter dataset. From the experiments performed, we can conclude that our pre-trained BERT models performs better than the baseline methods such as Logistic Regression and Support Vector Machines. One of the limitation of BERT is that it cannot handle long text sequences, however in our case we are using Twitter data which are mostly sequences of small texts making BERT a good choice here. This helps us to answer our research question concluding pre-trained BERT method is better than traditional baseline methods such as Logistic Regression and Support Vector Machines.

References

1. Bo Pang and Lillian Lee (2008): "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. https://doi.org/http://dx.doi.org/10.1561/1500000011
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". https://doi.org/https://arxiv.org/abs/1810.04805
3. Pak, Alexander Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10. *https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining*
4. Parikh, Ravi Movassate, Matin. (2009). Sentiment Analysis of User-Generated Twitter Updates using Various Classication Techniques. *https://www.researchgate.net/publication/242660794_Sentiment_Analysis_of_User-Generated_Twitter_Updates_using_Various_Classication_Techniques/citation/download*
5. Go, Alec Bhayani, Richa Huang, Lei. (2009). Twitter sentiment classification using distant supervision. Processing. 150. *https://www.researchgate.net/publication/228523135_Twitter_sentiment_classification_using_distant_supervision*
6. SemEval-2017 Task 4. https://doi.org/https://alt.qcri.org/semeval2017/task4/