ParsBERT Post-Training for Sentiment Analysis of Tweets Concerning Stock Market

Mohammadjalal Pouromid
Computer Science Dept.
Allameh Tabataba'i University
Tehran, Iran
jalalpouromid@gmail.com

Arman Yekkehkhani
Computer Science Dept.
Allameh Tabataba'i University
Tehran, Iran
arman.yekkehkhani@gmail.com

Mohammadreza Asghari Oskoei
Computer Science Dept.
Allameh Tabataba'i University
Tehran, Iran
oskoei@atu.ac.ir

Amin Aminimehr
Management Dept.
Ershad Damavand Institute of Higher Education
Tehran, Iran
aminaminimehr@outlook.com

2021 26th International Computer Conference, Computer Society of Iran (CSICC) | 978-1-6654-1241-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/CSICC52343.2021.9420569

***Abstract*—Social media has become a playground for users to share their ideas freely. Analyzing these data has become of special interest to authorities and consulting firms. They seek to choose right policies based on the insight acquired. Hence, sentiment analysis of data spread in social media has gained significant importance. There are two major approaches for sentiment analysis including lexicon-based and supervised methods. Among supervised methods, deep models have proven to be a better fit for the sentiment analysis task. Since, they are domain free and able to handle large volumes of data effectively. In particular, BERT's state of the art performance on various natural language processing tasks has encouraged us to use this network architecture for sentiment analysis. In this research, over 12000 Persian tweets including the stock market keyword have been crawled from twitter. They are labeled manually in three different categories of positive, neutral and negative. Then a pre-trained ParsBERT model has been fine-tuned on these data. Our model is evaluated on the test dataset and compared to its counterpart, lexicon-based method using Polyglot as its lexicon. Accuracy of 82 percent has been achieved by our proposed model surpassing its lexicon-based contender.**

***Index Terms*—sentiment analysis, deep learning, ParsBERT, twitter, stock market**

I. Introduction

Social media have become an indispensable part of our daily routine on which people are spending lots of their time to express what they feel. Expressing opinion freely with others through social media like twitter, facebook, etc is one of the reasons that encourages people to be an active member of it. The data available on social media has the characteristics of big data like velocity, heterogeneity and large-volume. Furthermore, it possesses a unique characteristic known as semantic, because it is generated manually and contains ambiguous subjective meaning [1]. Several challenges and opportunities in sentiment analysis have appeared for this unique characteristic.

Sentiment analysis also known as opinion mining is a way of monitoring crowd opinion about specific topics during a social crisis or predicting the satisfaction rate about a particular product or brand. A presidential candidate can arrange his advertising campaign or a company can fine-tune their production policies based on what people generally believe. In 2020 with the outbreak of the Corona virus and enforcement of quarantine regulations there was a possibility of economic collapse for every country. Therefore Iranian authorities decided to encourage people to invest in the stock market. At first, the index experienced a significant growth, and after about 5 months, it fell sharply causing a large amount of liquidity to be lost. Analyzing people's opinion on the subject of the stock market during this time can be of special interest.

Lexicon based methods are a way to find the sentiment score of a sentence based on sentiment orientation of the words existing in the lexicon instead of training on the data. These methods can be exploited to their full potential when the vocabulary is large enough. Polyglot[1] is a python package that provides a sentiment lexicon for Persian words. Lexicon based methods also have two disadvantages which are, the sentiment orientation of a word in a lexicon may be different from domain to domain and it is costly enough to search the sentiment orientation for every word in the lexicon [1]. It is reasonable to use a lexicon with similar semantics to our data in order to tackle the first disadvantage [2].

In supervised methods, sentiment of a text is predicted based on a contextually similar data labeled earlier. Unlike the conventional machine learning approaches, deep networks are more suitable for processing large datasets and also they are domain free. BERT [8], a fairly recent developed deep model for NLP tasks, is able to build a strong language model and be fine-tuned for specific tasks like sentiment analysis. State-of-the-art results of BERT on various tasks have proven uncontested potentials of this architecture.

[1]https://pypi.org/project/polyglot

978-1-6654-1241-4/21/$31.00 ©2021 IEEE

The following sections of the paper were organized as follows: In section 2, the related work to this article were discussed. Section 3 contains the structure of the model and the training data. Experiments such as collecting the data, preprocessing procedure and fine-tuning of the model were completely provided in Section 4. Results have been discussed in Section 5. In Section 6, the conclusion has been presented.

II. Related work

Nazan Ozturk and Serkan Ayvaz [3] extracted english and Turkish tweets about Syrian refugee crisis from twitter to find the opinion of Turkish people about it. They used lexicon based model to tag collected tweets. Ankita and Nabizath Saleenaa [4] have proposed an ensemble model for the classification task on four different datasets, and the accuracy of their proposed model outperformed other base machine learning algorithms.

Zhou et al. [5] have proposed a bi-directional LSTM model with 2D-pooling layer on Stanford database (Stanford Sentiment Treebank) in order to classify them; they reached the accuracy of 88.7%. Dos Santos and Gatti [6] have proposed a CNN model to classify the sentiment of some short texts by exploiting the character to sentence information. They named their model character convolutional neural network (ChCNN) which had two convolutional layers.

Farahani et al. [8] have pre-trained a monolingual model on a massive dataset of Persian text. The model is then fine-tuned to achieve state-of-the-art accuracy on different downstream tasks (e.g. Sentiment Analysis). They have reported the highest binary and multi-class F1 score of 71.11% and 92.13% on DeepSentiPers [9].

III. ParsBERT and Sentiment Analysis

In this section, a brief description of sentiment analysis task was presented and how it can be tackled with ParsBERT, a monolingual version of BERT trained specifically on Persian text corpora.

A. Sentiment Analysis

Sentiment Analysis also known as sentiment classification is a subtask of text classification whose purpose is to extract the implicit feeling contained in the text. Lexicon based and supervised based approaches are two major methods used for sentiment analysis. In lexicon based approach, the text is tagged based on its words orientation with respect to the lexicon. In this research, first the data was labeled by the authors, then one fifth of the data was used for the test phase and the rest for the training phase. The ParsBERT model was used as a supervised method. For lexicon based approach Polyglot was used as a vocabulary. Finally, the results were compared on the test dataset.

B. ParsBERT

Among pre-trained models, Transformer-based models such as BERT [7] have drawn lots of attention toward themselves in recent years. Popularity of these methods is basically due to their state-of-the-art performance. Furthermore, these models can be fine-tuned with minimal effort for specific tasks, such as sentiment analysis.

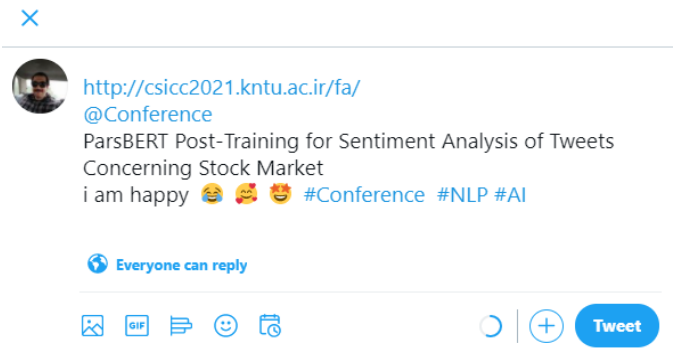

Fig. 1. Sample tweet with various features

ParsBERT [8] is a monolingual version of BERT trained on massive Persian text corpora, consisting of 3.9M documents, 73M sentences, and 1.3B words. This model has been able to outperform the multilingual BERT on a number of tasks, namely Sentiment Analysis(SA), Text Classification(TC), and Named Entity Recognition(NER). The model is based on BERT model architecture with a total number of 110M parameters. Particularly, it is founded on BERT BASE with the following configurations: 12 hidden layers, 12 attention heads and 768 hidden sizes.

IV. Experiments

A. Data Collection

Data were collected from January 1, 2020 to December 1, 2020 (the period when the stock market experienced both significant growth of index and its sharp decline) and subsequently were stored in a file formatted as excel. Every row of the data contains username and user_id of tweet's composer, text of the tweet and the posting date. A total of 12055 tweets were collected, none of which contains missing values. Twint[2] as a python package was used for extracting Persian tweets.

B. Preprocessing

Redundant features in tweets were pruned out in order to make the tweets suitable for learning procedure. Fig. 1 depicts a sample tweet with these features. Removing unwanted features is done as follows :

- Retweets started with "RT" were eliminated. external links and usernames proceed by were pruned out.
- Emojis were replaced by their meanings.
- Tweets with less than three words have been omitted.
- All punctuation and stop words were removed.
- Words were replaced with their stems.

[2]github.com/twintproject/twint

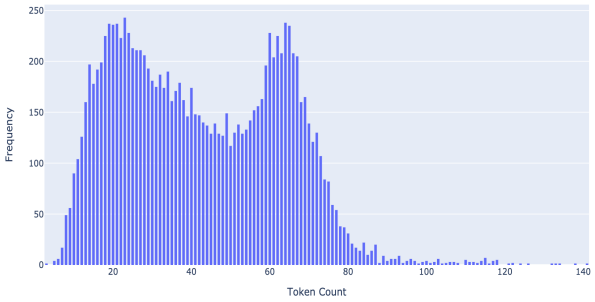

Fig. 2. Distribution of token counts within tweets

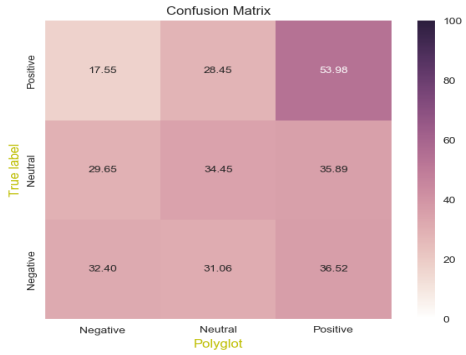

Fig. 3. Confusion matrix for test data tagged by a lexicon based method

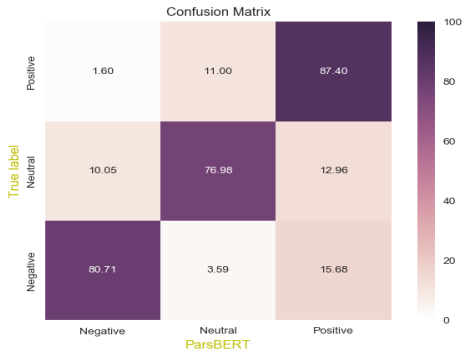

Fig. 4. Confusion matrix for test data tagged by fine-tuned ParsBERT

C. Fine-tuning

In this paper, ParsBERT was used as a pre-trained model and fine-tuned it on aforementioned data in a supervised manner. Processed input data were tokenized using the standard ParsBERT tokenizer, pre-trained alongside the main model, and fed to the model. Maximum input length was set to 128 tokens, based on the token counts distribution of the tweets illustrated in Fig. 2. Shorter sentences were padded with special token [PAD] to match the length of the largest one. Each input sequence starts with [CLS] token, the token whose embedding in the final layer would affect the classification result.

The model architecture had to be modified in order to comply with needs of the sentiment analysis task. This was done by adding a dropout layer and a fully connected layer, and a softmax layer to the end of the pre-trained model. Drop out rate was chosen to be 0.1, and the fully connected layer had 3 neurons corresponding to the number of output classes. The fully connected layer classified input tweets based on the first output of stacked encoders corresponding to [CLS] input token.

During training, cross entropy loss had to be minimized for a batch size of 24 with no clipping of gradient updates, which resulted in adjustment of encoder stack and fully connected layer weights. Training schedule included no warm-up steps. For optimization of the model, Adam optimizer was used with a learning rate of 2e-5, for 3 epochs.

V. Results

A. Compared Methods

Confusion matrix for the lexicon-based approach was shown in Fig. 3. For this approach, which used Polyglot as a lexicon, about 54% of tweets with positive labels were truly tagged positive and for the neutral and the negative classes this rate were about 33% and 35% respectively.

In this research, Precision, Recall and F-measure have been used for assessing the results. Precision is the fraction of relevant instances among the retrieved instances. recall is the fraction of the total amount of relevant instances that were actually retrieved and F-measure is the harmonic mean of precision and recall. The best value for the precision criterion was for the negative class with 48.7%. For recall and f-measure its best values were for positive and negative classes by 53.99% and 38.91% respectively. On average, for the three classes, an accuracy of 38.27% has been achieved for the precision, 40.28% for the recall, and 36.5% for the f-measure.

Confusion matrix for the supervised approach was shown in Fig. 4. For this approach, which used ParBERT as a classifier, about 87.4% of tweets with positive labels were truly tagged positive and for the neutral and the negative classes this rate were about 76.98% and 80.71% respectively.

The best value for the precision criterion was for the positive class with 89%. For recall and f-measure its best values were for the positive class by 87.4% and 88.2% respectively. On average, for the three classes, an accuracy of 82.04% has been achieved for the precision, 82.6% for the recall, and 81.83% for the f-measure.

B. Trend Overview

Word-cloud is an overview of the most frequent words present in preprocessed tweets. Corresponding word-cloud is

TABLE I
Results of lexicon based method

Class	Precision	Recall	F-measure
Negative	48.7	32.41	38.91
Neutral	43.04	34.45	38.27
Positive	23.07	53.99	32.32
Average	38.27	40.28	36.5

Fig. 5. Word-cloud of the processed tweets

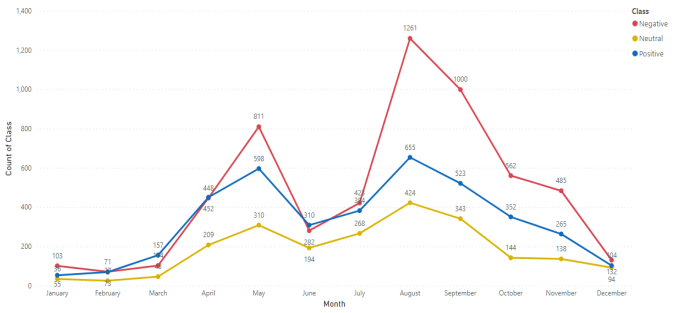

Fig. 6. Overall trend of the tweets frequencies with respect to their sentiment classes

illustrated in Fig. 5.

Fig. 6 shows the overall trend of the tweets frequencies with respect to their sentiment classes. The red line, the blue line and the yellow line belong to the negative, the positive and the neutral classes respectively. As it can be seen in Fig5, the positive tweets had the higher hand in number until May due to the market growth. Afterwards, during the first pull-back of the index, the number of negative tweets rose. Termination of pull-back led to an increase in the number of positive tweets. In August, sharp fall of the index resulted in a drastic rise of the negative to positive tweets ratio. In December, tweets were distributed in three classes equally, by the market pacificity.

VI. Conclusion

Persian tweets provide a great source of data representing a realistic reflection of people's ideas, deeds, and feelings. Despite their abundance, tweets are mostly written in informal and sarcastic language. Sentiment analysis of these data with lexicon-based methods has been shown to yield poor performance. However, deep models based on BERT architecture are able to capture complex semantics well, and outperform their counterparts. These deep models can be used to provide reliable analysis for consulting firms.

TABLE II
Results of fine-tuned ParsBERT classifier

Class	Precision	Recall	F-measure
Negative	84.12	80.71	82.37
Neutral	72.98	76.98	74.92
Positive	89.03	87.4	88.2
Average	82.04	82.6	81.83

References

[1] Behera, Ranjan Kumar, Monalisa Jena, Santanu Kumar Rath, and Sanjay Misra. "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data." Information Processing & Management 58, no. 1: 102435.

[2] Cambria, Erik. "An introduction to concept-level sentiment analysis." In Mexican international conference on artificial intelligence, pp. 478-483. Springer, Berlin, Heidelberg, 2013.

[3] Ozturk, Nazan, and Serkan Ayvaz. "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis." Telematics and Informatics 35, no. 1 (2018): 136-147.

[4] Saleena, Nabizath. "An ensemble classification system for twitter sentiment analysis." Procedia Computer Science 132 (2018): 937-946.

[5] Zhou, Peng, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling." arXiv preprint arXiv:1611.06639 (2016).

[6] Dos Santos, Cicero, and Maira Gatti. "Deep convolutional neural networks for sentiment analysis of short texts." In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 69-78. 2014.

[7] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[8] Farahani, Mehrdad, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. "ParsBERT: Transformer-based Model for Persian Language Understanding." arXiv preprint arXiv: 2005.12515 (2020).

[9] Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. ArXiv, abs/2004.05328, 2020.