Ternary Sentiment Classification of Airline Passengers' Twitter Text Based on BERT

Yungao Xie*, Hong Wen and Qing Yang
School of Computer Science and Engineering, Hunan University of Science and Technology, Hunan, China

*Corresponding author email: 18010502005@mail.hnust.edu.cn

Abstract. Traditional word vector generation model cannot solve the polysemy problem of word representation in Twitter text, so, a method is presented here, firstly, by using Bidirectional Encoder Representations from Transformers (BERT), the semantic feature vector of the text can be obtained, and then, the feature vector is inputted into the Softmax classifier to implement the sentiment classification of Twitter text. The experimental datasets are source from passengers' Twitter comments of USA six airlines, the sentiment classification model based on Embeddings from Language Model (ELMo) as the experimental control group. The experimental results indicate that the proposed model is advantage over the experimental control group by using F1-score as the evaluation index.

Keywords: BERT; Sentiment classification; Airline passengers; Twitter texts; Softmax.

1. Introduction

With the rapid popularization of social networking applications based on user-generated content\(^1\) (UGC) such as Facebook, Weibo and Twitter, an abundant of short texts updated every day, these short texts published on social media contain the sentiment tendency of users, which include user preferences, user attitudes and user values. By sentiment classification of passengers' Twitter texts, passengers' sentiment tendencies can be automatically acquired, then airlines can dynamically adjust service strategies, which are conducive to the precision marketing.

The Twitter text has the characteristics of concise words and networked expressions, there are a lot of polysemous words in the Twitter text, for example, "airline" can express two meanings of "shipping line" and "airways" in different contexts, and "book" can also express two meanings of "text paper" and "reserve". Traditional word vector generation models (BoW, TF-IDF\(^2\), Word2vec\(^3\), GloVe\(^4\), etc.) cannot be solved the polyseymy problem of word representation, which became an anxious problem in the research field of sentiment classification of UGC texts such as Twitter texts\(^5\).\(^6\).\(^7\).\(^8\).

To solve the polyseymy problem of word representation, ELMo\(^9\) has been proposed by ME Peters in 2018, but feature extractor of ELMo is bidirectional LSTM, which cannot be used in parallel computing. Transformer\(^10\) has been proposed by Google. By using Transformer, all the words in the sequence can be processed, and the context relationship can be extracted by using the self-attention mechanism. GPT\(^11\) has been proposed by OpenAI, feature extractor of GPT replaces bidirectional LSTM\(^12\) with unidirection Transformer, which can be used to obtain semantic information of sentences from left to right. BERT\(^13\) has been proposed by Jacob Devlin in October 2018, BERT and ELMo have the same architecture, except that feature extractor of BERT replaces bidirectional LSTM with bidirection Transformer. By using BERT, the performance of the 11 NLP tasks was excellent, which make BERT...
become the most popular word vector generation model. To solve the polysemy problem of word representation, we proposed the ternary sentiment classification model based on BERT.

2. The Ternary Sentiment Classification Model Based on BERT

We proposed the ternary sentiment classification model based on BERT, and the structure of the proposed model is shown in figure 1, firstly we process the airline passengers' Twitter text, then, by using BERT, transform texts into the semantic feature vector matrix which can be recognized by machine, and finally input feature vector into Softmax classifier to implement ternary sentiment classification.

2.1. BERT Model

BERT model is used to extract the feature of text to solve the polysemy problem of word vector representation of Twitter text. The structure of BERT[13] is shown below:

The language model of Bert is built by bidirectional Transformer. Transformer is composed of encoder-Decoder and BERT is composed of the encoder layer of Transformer. The core module of Transformer encoding layer is the self-attention mechanism. Recurrent Neural Network[14] cannot be used in parallel computing and CNN[15] cannot be used in extracting global information at once, Transformer models text sentences entirely by using self-attention mechanism[10]. The calculation formula of Attention is as follows:
\[ \text{Attention}(Q, K, V) = \text{soft}(\frac{QK^T}{\sqrt{d_k}})V \]  

\( Q, K, V \) are vector matrix, \( d_k \) is the dimension of the input vector. The core idea of self-Attention mechanism is to calculate the relationship between each word in the sentences, using the relationship between the weights to adjust each word can get a new expression, this new expression not only contains the relationship between other words but also the word itself. After the training, as shown in figure 1, \( T_1, T_2, \ldots, T_N \) is the semantic representation of each word, and \( C \) is the semantic representation of the whole sentence. In this research, the feature vector \( C \) which can be represented the whole sentence semantics is used as the input of the classifier.

2.2. Softmax Classifier

Softmax model can be used to construct the classifier. Softmax model is the extension of Logistic regression model on multivariate classification. When the classification number is 2, Softmax model rollback to Logistic regression model. The calculation formula of the sentiment classification model is as follows:

\[ P(y = j | x, W_0, b_0) = \text{soft}(W_0 \cdot C + b_0) \]  

\( W_0 \) is weight coefficient matrix, \( b_0 \) is bias matrix, and each input \( x \) in the model has a probability \( P(y = j | x) \) for each category, \( j = 1, 2, \ldots, k \), the category with the highest probability is the category to which \( x \) belongs.

3. Experiments

3.1. Datasets

The passengers' Twitter texts of six major USA airlines, collected by Crowdflower in February 2015, are used as the experimental datasets. The distribution of the datasets is shown in table 1.

|                | American | Delta | Southwest | United | USAir | VA |
|----------------|----------|-------|-----------|--------|-------|----|
| negative       | 1960     | 955   | 1186      | 2633   | 2263  | 181|
| neutral        | 463      | 723   | 664       | 697    | 381   | 171|
| positive       | 336      | 544   | 570       | 492    | 269   | 152|

3.2. Evaluation Index of Ternary Sentiment Classification

\( F1 - \text{score}_{\text{micro}} \), \( F1 - \text{score}_{\text{macro}} \) and \( F1 - \text{score}_{\text{weighted}} \) are adopted as the experimental evaluation indexes of ternary sentiment classification. \( F1 - \text{score}\) can be regarded as the harmonic average value of recall and precision. There are three different calculation methods of \( F1 - \text{score}\).

The micro average is the arithmetic average of the performance index of each sample, which is susceptible to the influence of the large category:

\[ F1 - \text{score}_{\text{micro}} = \frac{2 \cdot \text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \]  

The macro average is the arithmetic mean of the performance index of each category, which is susceptible to the influence of the small category:

\[ F1 - \text{score}_{\text{macro}} = \frac{2 \cdot \text{Precision}_{\text{macro}} \cdot \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}} \]
The weighted average method is obtained by multiplying and adding the proportion of all kinds of samples in all sample as the weight and all kinds of performance indexes:

$$F1-score_{weighted} = \frac{2 \times Precision_{weighted} \times Recall_{weighted}}{Precision_{weighted} + Recall_{weighted}}$$  \hfill (5)

### 3.3. The Procedure of Experimentation

TensorFlow is adopted to implement the modeling of ternary sentiment classification model based on BERT, and the Python programming language is used for experimental analysis on Kaggle platform. Firstly, the pre-trained BERT model file is loaded, then the fine-tuning operation is carried out. By using BERT, the semantic feature vector of the processed datasets can be obtained, and then, the feature vector is inputted into the Softmax classifier to implement the sentiment classification of the Twitter text. We set the sentiment classification model based on ELMo as the experimental control group, the feature vector of text can be extracted by using ELMo, then vector is inputted into Softmax classifier to implement the sentiment classification.

### 3.4. Experimental Results and Analysis

We use the proposed model to implement the sentiment classification of Twitter text, the experimental datasets are passengers’ Twitter comments text of USA six airlines, the evaluation index of ternary sentiment classification is $F1-score_{micro}$, $F1-score_{macro}$ and $F1-score_{weighted}$.

**Figure 3.** The comparison of $F1-score_{micro}$

**Figure 4.** The comparison of $F1-score_{macro}$

**Figure 5.** The comparison of $F1-score_{weighted}$

Figures 3-5 display the comparison of $F1-score_{micro}$, $F1-score_{macro}$ and $F1-score_{weighted}$ which were predicted by using the proposed model and the experimental control group on the six datasets. The regularity of $F1-score_{micro} > F1-score_{macro} > F1-score_{weighted}$ is presented by using the other five experimental datasets except for the VA dataset, the regularity is caused by the unbalanced distribution of sentiment polarity in the experimental datasets. The proportion of the text label as the negative category is relatively large, and it has more data features that can be extracted, which is conducive to the classification prediction of negative category. Similarly, the proportion of the text label as the positive category is small, which is not conducive to the prediction of the positive category.
micro F score value is easily affected by the negative category, so the value is relatively large.
macro F score value is easily affected by the positive category, so the value is relatively small.
weighted F score that the sample quantity calculated as weight is not affected by the sample distribution,
so the value of weighted F score is in the middle, which can more effectively evaluate the classification prediction effect of the model.

Figures 3-5 show that the classification performance of the proposed model, by using micro F score as the evaluation index, is superior to the model based on ELMo on six datasets. By using macro F score as the evaluation index, the VA dataset have the biggest difference, the proposed is 4.95% higher than the model based on ELMo. By using weighted F score as the evaluation index, the biggest difference is the United dataset, which is 10.98% higher. By using weighted F score as the evaluation index, the biggest difference is also the United dataset, which is 5.74% higher. In conclusion, the semantic information of airline passengers' Twitter texts can be excellently extracted by using BERT, which can promote the sentiment classification of Twitter texts.

4. Conclusion
In the field of airline passengers' Twitter text sentiment classification, traditional word vector generation model cannot solve the polysemy problem of word representation, so, we propose ternary sentiment classification model based on BERT. The passengers' Twitter comments text of six major U.S airlines are used as the experimental datasets. The sentiment classification model based on ELMo is used as the experimental control group. The comparative experiments indicate that the proposed model is advantage over the model based on ELMo by using micro F score, macro F score and weighted F score as the evaluation index, which verifies the excellent performance of BERT in the field of word representation.
The work of this paper has a certain significance for the research of UGC text sentiment classification. UGC text has the characteristics of concise words and networked expressions, words of UGC text have rich connotations. By using BERT, the semantic feature vector of the UGC text can be obtained, and the polysemy problem in different contexts can be solved, which provides new solutions and research ideas for UGC text sentiment classification. The method also has some disadvantages, such as running BERT on the computer has high computational complexity, we try to address those issues in the future.

References
[1] Reyes-Menendez 2020 Using discourse analysis in UGC. International Journal of Information Management, p 54
[2] Jones K S 1973 Index term weighting. Information Storage & Retrieval, 9(11) pp 619-633
[3] Mikolov T 2013 Efficient Estimation of Word Representations in Vector Space. Computer Science, (1) p 256
[4] Pennington J 2014 Glove: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing
[5] Asgarnezhad R 2019 A High-Performance Model based on Ensembles for Twitter Sentiment Classification. Journal of Electrical and Computer Engineering Innovations (JECEI), 8(1) pp 41-52
[6] Alharbi A S M 2019 Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54 pp 50-61
[7] Naseem 2020 Transformer based deep intelligent contextual embedding for twitter sentiment analysis. Future Generation Computer Systems, 113 pp 58-69
[8] YuWen Lyu 2020 Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China’s social media Weibo. Children and Youth Services Review
[9] Peters M E 2018 Deep contextualized word representations. arXiv
[10] Vaswani A 2017 Attention Is All You Need. arXiv
[11] Radford 2018 Improving language understanding by generative pre-training
[12] Hochreiter S 1997 Long short-term memory. Neural computation, 9(8) pp 1735-1780
[13] Devlin J 2018 Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv
[14] Rumelhart D E 1986 Learning representations by back-propagating errors. Nature, pp 533-536
[15] Lecun Y 1998 Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11) pp 2278-2324