Discriminative Features Fusion with BERT for Social Sentiment Analysis

Duy-Duc Le Nguyen[1], Yen-Chun Huang[1] , and Yung-Chun Chang[1,2]

Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan
{m946108006, m946108007, changyc}@tmu.edu.tw

Abstract. The need for sentiment analysis in social networks is increasing. In recent years, many studies have shifted from author sentiment research to reader sentiment research. However, the use of words that hinders sentiment analysis is very diverse. In this paper, we provide a model that combines the latest and most recent contextual text embedding technology and feature selection to more accurately detect the emotional intent of an article. We named it DF2BERT(Discriminative Features Fusion with Bert), and extensively applied datasets in different languages and different text classification tasks to validate our method, and compared it with several well-known approaches. Experimental results show that our model can effectively predict sentiment behind the text which outperform comparisons.

Keywords: Natural Language Processing, Sentiment Analysis, Text Representation, Feature Fusion, Deep Neural Network

1 Introduction

At present, big data analysis technology has rapidly developed. According to previous research, most of the unstructured data are currently buried in text data, and they are mainly distributed on the Internet. Today, the online community and news media have become the main field of information acquisition and transmission for modern people. For instance, if we want to buy new sneakers, we would collect reviews of products from the online community and weigh our decision based on others suggestions.

In addition, sentiment analysis research is getting more and more attention with an attempt to obtain trends in public by mining opinions that are subjective statements that reflect people's sentiments or perceptions about topics [11]. Thanks to the optimization of natural language processing (NLP) technology. It helps us to accurately analyze the sentiment and opinions of massive texts. While previous researches on emotions mainly focused on detecting the emotions that the authors of the documents were expressing. It is worthy of note that the reader-emotions, in some aspects, differ from that of the authors and may be even more complex [6, 12]. For instance, a news article with the title "The price of crude oil will rise 0.5 % next week" is just objectively reporting an event without any emotion, but it may invoke emotions like angry or worried in its readers. Furthermore, it is possible to get more sponsorship opportunities from the company or manufacturer if the articles describing a certain product are able to promote greater emotional resonance in the readers.

In this paper, we proposed the method which is discriminative features fusion with BERT (DF2BERT) for social sentiment analysis. We first extract the discriminative features from raw text. Then, we integrate the features into BERT for sentiment classification. Our experiments demonstrate that DF2BERT can achieve a higher performance than other well-known methods of text categorization on two different sentiment categorization tasks. Furthermore, the proposed method embraces the advantages of both feature-based and deep learning-based approaches. First of all, as shown in our experiments, it is language-independent. Second, compared to feature-based systems, DF2BERT is automatic and scalable. We believe DF2BERT points to a promising direction for many natural language applications.

2 Discriminative features fusion with BERT model

We model sentiment analysis as a classification problem, and define both task as the following. Let $W = \{w_1, w_2, \ldots, w_n\}$ be a set of words, $D = \{d_1, d_2, \ldots, d_m\}$ be a set of documents, and $C = \{c_1, c_2, \ldots, c_k\}$ be a set of categories. Each document d_x is a set of words W_{d_x} such that $W_{d_x} \subseteq W$. The goal of this task is to decide the most appropriate sentiment s_i for a document d_j, although one or more sentiments can be associated with a document. In this paper, we first learn the keyword of sentiments from raw text. The learned keywords are converted to discriminative feature vectors (DFV) for representing sentiments. Afterward, we further integrate the DFV with BERT for the tasks of sentiment analysis. We will describe the proposed method in the following sections.

Fig. 1 illustrates an overview of the DF2BERT model. E is the embedding representation and T is the final output of BERT architecture. There are 12 layers in the original BERT model where each token will have 12 intermediate representations. Every sequence takes the special classification token $[CLS]$ as the first token which is followed by the WordPiece tokens. For classification tasks, $[SEP]$, the separator token, can be ignored. The maximum sequence of length of the input are 512 tokens. The final hidden state of the $[CLS]$ token is taken as a fixed pooled representation of the input sequence. This is a vector with 768 size represent the whole input sequence for classification tasks. Since BERT has been pre-trained for Masked Language Modeling and Next Sentence Prediction, it also takes masking and next sentence representation as inputs. But classification tasks do not need to consider about masking and next sentence predicting, masking representation is an array of 1 value and next sentence representation is a 0 one.

BERT for sequence classification takes one segments (sequences of tokens) as input. The segment is re-presented as a single input sequence to BERT with an special tokens: $[CLS], x_1, \ldots, x_N$. N need to be less or equal to L which is a parameter that controls the maximum sequence length during training. At the output, the $[CLS]$ representation is fed into an output layer for the sentiment

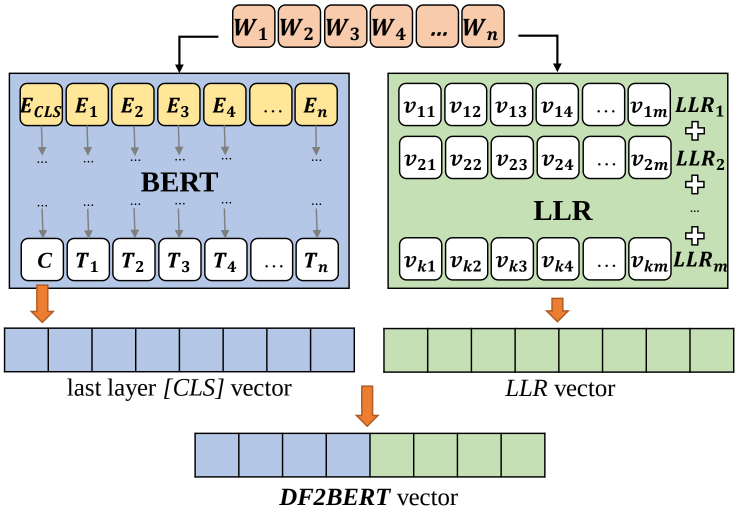

W_1 W_2 W_3 W_4 ... W_n
E_{CLS} E_1 E_2 E_3 E_4 … E_n
BERT
C T_1 T_2 T_3 T_4 … T_n
v_{11} v_{12} v_{13} v_{14} … v_{1m} LLR_1
v_{21} v_{22} v_{23} v_{24} … v_{2m} LLR_2
LLR
v_{k1} v_{k2} v_{k3} v_{k4} … v_{km} LLR_m
last layer *[CLS]* vector
LLR vector
DF2BERT vector

Fig. 1. Architecture of the discriminative feature fusion with BERT model(DF2BERT).

analysis. Adam [4] is the optimization of BERT with the following parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{(-5)}$ and $L_2 = 0.01$. BERT trains with 0.1 dropout rate on all layers and attention weights with activation function as GELU.

Furthermore, it is common to reduce the number of terms used in text classification. To increase classification accuracy by eliminating noise features. A noise feature is kind of that when added to the document representation, increase the classification error on new data. For instance, the Bernoulli model is sensitive to noise features. Moreover, sentiment lexicons are important information for identifying sentiment behind the text. In light of this, we use the log likelihood ratio (LLR) [10], an effective feature selection method, to learn a set of sentiment-specific lexicons. Given a training dataset, we first obtain four frequencies $k = N(L \wedge S)$, $l = N(L \wedge \neg S)$, $m = N(\neg L \wedge S)$, and $n = N(\neg L \wedge \neg S)$, in which $N(L \wedge S)$ denotes the number of documents that contain l and belong to sentiment S, $N(L \wedge \neg S)$ denotes the number of documents that contain L but does not belong to sentiment S, and so on. LLR employs Equation to calculate the likelihood of the assumption that the occurrence of a lexicon l in sentiment T is not random.

$$LLR(L,S) = 2log\left(\frac{p(L|S)^k(1-p(L|S))^m p(L|\neg S)^l(1-p(L|\neg S))^n}{p(L)^{k+l}(1-p(L))^{m+n}}\right) \quad (1)$$

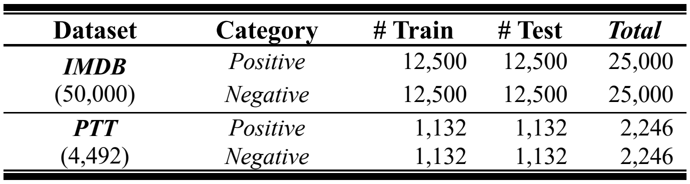
In (1), probabilities $p(L)$, $p(L|S)$, and $p(L| \wedge S)$ are approximated using maximum likelihood estimation. A lexicon with a large LLR value is closely associated with the sentiment. We rank the lexicons in the training data based on their LLR values and select lexicons with high LLR values to compile a sentiment lexicon list.

3 Performance Evaluation

3.1 Dataset and Setting

In order to examine the flexibility and robustness of the proposed method, we use publicly available datasets as well as those collected on our own for evaluation. Table 1 shows the descriptive statistics of three datasets used in our experiments. We tested the proposed method on a sentiment classification of IMDB movie review dataset [9] (denoted as IMDB). This is a balance dataset which containing the same proportion of positive and negative in 25,000 movie reviews for training, and 25,000 for testing. In order to evaluate the reliability, we tested DF2BERT on a small Chinese movie review corpus [1], which collected from an electronic bulletin board system that provides a social platform for academic purposes (denoted as PTT).

Table 1. Descriptive statistics of three datasets used in text categorization experiments. Total amount of documents in each dataset is listed in parentheses.



Dataset	Category	# Train	# Test	*Total*
IMDB	*Positive*	12,500	12,500	25,000
(50,000)	*Negative*	12,500	12,500	25,000
PTT	*Positive*	1,132	1,132	2,246
(4,492)	*Negative*	1,132	1,132	2,246

To evaluate the effectiveness of the compared systems, we adopt the convention of classification accuracy for sentiment analysis [5], and use macro-average for the estimation of overall performance. These measures were defined based on a contingency table of predictions for a target sentiment S_k. We use bert-base-uncased as the pre-trained model for both BERT model and BERT tokenizer from transformer package. We re-implement BERT with RoBERTa setting [8]. We primarily follow the original BERT optimization hyper-parameters given in Section 3, except β_2 is set as 0.98 to improve stability when training. The maximum sequence length is 512 tokens where padding or truncating at the end of segment. We run BERT model in a single GPU RTX 2080 Ti. Since it only have 11GB GPU RAM, 6 sequences is trained per batch As an out, the last hidden layer of $[CLS]$ which represent a whole sentence will concatenate with the output layer of feature extraction step. Here, we extract top 70 features affected to each label. During training time, we have found that 70 extracted features achieved better performance compare with 30, 50, 100, 150 and 200 features.

3.2 Result and Discussion

The purpose of the experiments was to examine the effectiveness of the DF2BERT method in capturing discriminative lexicon information and representing it for a text classification model for sentiment analysis task in different language corpora. A comprehensive performance evaluation of the DF2BERT with 10 well-known text classification methods is provided. The first is a keyword-based model which adopts TF-IDF term weighting and is trained by SVM (denoted as SVM) with linear kernel. Another is the random forest approach which consist a large number of relatively uncorrelated decision trees operating as an ensemble (denoted as RF). Next, is a gradient boosting decision tree that integrates multiple learners for classification problems (denoted as XGBoost) Two well-known deep learning-based text classification approaches were also included: convolutional neural network for text classification (denoted as TextCNN) [3] and recurrent neural network with long short-term memory (denoted as LSTM) [7]. Moreover, comparing DF2BERT with BERT enables us to verify the contribution of the proposed discriminative feature fusion techniques. To serve as a standard for comparison, we also included the results of Naive Bayes (denoted as NB), Decision Tree (denoted as DT), Logistic Regression (denoted as LR), and k-Nearest Neighbors (denoted as KNN) as baselines. The performances of sentiment analysis and topic detection systems are listed in Table 2.

Table 2. Comparison of the performance of 11 classification systems of IMDB and PTT corpus. Bold numbers indicate the best performance in the category.

Method	**IMDB**	**PTT**
	Accuracy (%)	
NB	82.96	80.04
DT	70.37	70.01
LR	88.32	84.41
KNN	66.15	70.98
SVM	88.26	87.81
RF	72.95	74.29
XGBoost	81.04	81.89
TextCNN	82.61	53.45
LSTM	85.40	78.05
BERT	93.21	87.38
DF2BERT	**93.48**	**88.18**

As the results, most of machine learning-based approaches can achieve about 80% classification accuracy. For baseline models, regarding the binary-class dataset, we can see that in Logistic Regression and SVM, the performance is better than the tree structure model, and the performance results of Naive Bayes and XG-Boost are closely behind. Text classification contains many features and is mostly linearly separable [2]. Logistic Regression is better than the tree structure model in dealing with linear separable classification problems. It is more effective in dealing with outliers in the data and avoids over-fitting the training set to the test set. It is worth noting that the SVM can achieve good performances. This can be partially explained by using the statistical learning theory argument of SVM in. If there exists a linear separator w with a small 2-norm that separates in-class data from out-of-class data with a relative large margin, an SVM can perform well.

In general, deep learning-based methods can achieve about 85% classification accuracy on IMDB dataset. However, deep learning-based methods inferior in PTT dataset. This may be due to the insufficient amount of data in the PTT dataset itself, which is difficult to reflect the power of deep learning. Moreover, the limitation of the neural network-based methods is that it cannot take the order of the inputs into account like recurrent neural network. If the first and last token are the same word, it will treat those tokens exactly the same. By contrast, BERT overcomes this problem through learning the context and semantic information. Consequently, it performs better than other machine learning and deep learning methods. Notably, DF2BERT achieves the best performance with approximating 94% accuracy since we further integrate the discriminative features into BERT model. This is because the proposed method can extract useful words and terms in each class so that the vector represents a sequence generated by BERT with the extra LLR vector at the end will orient to the right class.

To summarize, the proposed discriminative feature fusioin with BERT approach successfully integrates the syntactic, semantic, and content information in text to recognize sentiment behind document. Hence, it achieves the best classification accuracy, as shown in Table 2.

4 Conclusion

In this paper, we propose a new model to improve text embedding results, combining feature selection algorithms with the latest natural language processing method. The model was trained to deal with cross-language classification in English and Chinese data. Our model is applied to a total of four different datasets with various classification task: binary, and multiple class classification. The results show that our method is very effective in binary and multiclass classification, especially with big dataset such as IMDB movie review and CND news. For small dataset like PTT movie review, DF2BERT model still remain leading place among all methods.

In the future, we plan to refine DF2BERT and employ it to other NLP applications. We will also investigate the syntactic dependency information in social media to incorporate further syntactic and semantic information into the BERT structure.

References

1. Chu, C.H., Wang, C.A., Chang, Y.C., Wu, Y.W., Hsieh, Y.L., Hsu, W.L.: Sentiment analysis on chinese movie review with distributed keyword vector representation. In: 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI). pp. 84–89. IEEE (2016)
2. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. pp. 137–142. Springer (1998)
3. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Lin, K.H.Y., Yang, C., Chen, H.H.: What emotions do news articles trigger in their readers? In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 733–734. Citeseer (2007)
6. Lin, K.H.Y., Yang, C., Chen, H.H.: Emotion classification of online news articles from the reader's perspective. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. pp. 220–226. IEEE Computer Society (2008)
7. Liu, P., Qiu, X., Huang, X.: Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101 (2016)
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
9. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. pp. 142–150. Association for Computational Linguistics (2011)
10. Manning, C., Schütze, H.: Lexical acquisition. In: Foundations of statistical natural language processing, vol. 999, pp. 296–305. MIT press (1999)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)
12. Tang, Y.j., Chen, H.H.: Mining sentiment words from microblogs for predicting writer-reader emotion transition. In: LREC. pp. 1226–1229 (2012)