Sentiment Classification of Reviews Based on BiGRU Neural Network and Fine-grained Attention

IOP Conf. Series: Journal of Physics: Conf. Series **1229** (2019) 012064 doi:10.1088/1742-6596/1229/1/012064

Sentiment Classification of Reviews Based on BiGRU Neural Network and Fine-grained Attention

Xuanzhen Feng[1,*] and Xiaohong Liu[1]

[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

[*]Corresponding author: fengxuanzhen@bupt.edu.cn

Abstract. Text sentiment analysis is part and parcel of natural language processing. The task of sentiment classification is actually the process of feature extraction through models. The comment text of commodities is very different from the ordinary text. The comment text has no fixed grammar and writing format and the sentiment feature information is scattered in various places of text. Due to these factors, model learning of sentiment classification is becoming increasingly complex. The paper aims at establishing a fine-grained feature extraction model based on BiGRU and attention. Firstly, the vocabulary is vectorized by means of the skip-gram model. Then, according to the pre-trained word vector, the sentiment words list can be reached and noise filtering would be conducted by Naive Bayes algorithm. Finally, the model extracts features using BiGRU and fine-grained attentions. Based on the hypothesis that a long review may lead to feature differentiation, a fine-grained attention model is proposed. In this model, the attention layer is design to focus on the feature in different level such as word level, sentence level and paragraph level. This paper validate the proposed model on two sentiment corpus JD reviews and IMDB. Empirical results show that the FGAtten-BiGRU model achieves state of the art results on sentiment analysis tasks.

1. Introduction

There are some characteristics for reviews, such as a certain length, limited vocabulary and no strict grammatical rules [1]. The mainly sentiment classification model for text is based on sentiment polarity lexicons and traditional machine learning which include the construction of sentiment resource, extraction of feature information sentiment classification, quality analysis and so on [2]. Sentiment classification model based on sentiment polarity lexicons is commonly used, but the existing sentiment vocabulary lexicons is limited. For example, the model based on lexicons can not recognize a newly-coined word or a buzzword effectively. Manek and Shenoy used traditional machine learning algorithms to analyze the sentiment of reviews [3]. They mainly compared Naive Bayes, ME and SVM in accuracy and F score value [3]. The results showed that SVM had the best classification effect. With the development of deep learning research, the deep neural network presents outstanding performance in natural language processing [3]. Kim proposed using Convolutional Neural Network (CNN) to solve sentiment classification problem and achieved good results [4]. Santos used deep convolution neural network to analyze sentiment in short texts [5]. Irsoy used recurrent neural networks (RNN) to model sentences, and LSTM (long short-term memory), which is a model of recurrent neural networks, was also proved effective to solve sentiment analysis problems [5]. Attention model, proposed by Bahdanau, was used in machine translation initially, then attention's

different variants were widely used in NLP [6]. Qu Zhaowei and Wang Yuan proposed a sentiment analysis model based on hierarchical attention network, which improved by five percentage points compared with the traditional recurrent neural network [7].

This paper designs a model based on hierarchical attention and Bi-GRU network for sentiment classification of reviews. We find that the variable length of input may lead to sparse matrix and obscure features. In order to solve this problem, we proposed a fine-grained attention mechanism, which divides attentions into different granularities and calculated them synthetically. By comparing the models from different types of comment sets, we can conclude that our method can change dimension flexibly and improves the accuracy of feature extraction.

2. Related work

2.1. Data Acquisition and Preprocessing

The construction of data set is a fundamental part for model [8]. Due to the fact that review data of Jingdong is filtered, which can help us save some data processing work, we choose Jingdong Mall as the data source in this paper. In order to guarantee the uniform distribution of data, we crawled the comment data in different commodity types. In this paper, we crawled JingDong comments from mouse pages, televisions pages, mobile phones pages and laptops pages as data sources and chose the same number of positive and negative records for each commodity item. In addition, in order to verify the robustness of the model in sentiment analysis tasks, we also use IMDB data set as input source.

2.2. Word Embedding

One of the most important points in natural language processing is to transfer human-recognizable characters to computer-recognizable symbols [8]. Google proposed two architectures of word vectors, CBOW model and Skip-gram model [9]. They also developed Word2vec tools that can be used to compute billions of corpus data, which greatly promoted the development of natural language processing [10]. CBOW model predicts the current word from a window of surrounding context words, Skip-gram model uses the current word to predict the surrounding window of context words [10]. Compared with word bag model, the word vector trained by neural network is more reasonable and effective.

3. Model design

3.1 Noise reduction based on fine-grained word vector

Word2vec model is used for word vector generator in this paper. Although the preprocessing can drop off the noise data, there still exist a large number of unreasonable records in the data set. One of the common problems is that users input a positive polarity data with a negative polarity type, which will lead to logical errors [11]. In this paper, fine-grained word vectors are used to correct these logic errors to some extent. Although the noise cannot be completely avoided, it can be minimized to a great degree. Sixteen similar words are extracted from the word vectors according to the words of "好" ("good") and "差" ("bad") respectively as feature words, and the frequency of these words appearing on different class is counted as probability. Then the conditional probability of the data is calculated by Naive Bayes method, and the polarity of the record is reversed only when the result value exceeds the threshold value:

$$\mathrm{y}_{\mathrm{pos}} = \mathrm{P}(y_0 \mid x_1, x_2, ...x_{15}) = P(Y = y_0)\prod_i P(X^{(i)} = x^{(i)} \mid Y = y_0) \tag{1}$$

$$\mathrm{y}_{\mathrm{neg}} = \mathrm{P}(y_1 \mid k_1, k_2, ...k_{15}) = P(Y = y_1)\prod_i P(K^{(i)} = k^{(i)} \mid Y = y_1) \tag{2}$$

$$\text{change Label to positive} = if\ (y_{pos} > threshold_{\mathrm{pos}}) \tag{3}$$

$$\text{change Label to negative} = if\ (y_{neg} > threshold_{neg}) \tag{4}$$

Matrix sparsity is often encountered when comments are inputted. The reason for this problem lies in the variable size of text, which results in a large number of zero padding. Therefore, we divides paragraph into pieces sentences based on bucket.

3.2 Fine-grained Attention Fusion Network Model

The attention model is similar to the way humans allocate attention. It does not capture all the details of a sentence, but focus on the key points. Based on the model proposed by Zhou et al [12], our model designs a fine-grained attention which combines attention model and hierarchical attention model to learn the key points. For long sentences, we inputs previously segmented data into different level attentions. The purpose of hierarchical processing of sentences is to extract more detailed features, but when sentences are short to a certain extent, short data is not suitable for hierarchical attention model, therefore, coarse-grained model is needed. So in the model, we adopt two neural network channel to extract the fine-grained feature. The schematic diagram of the model as figure 1:

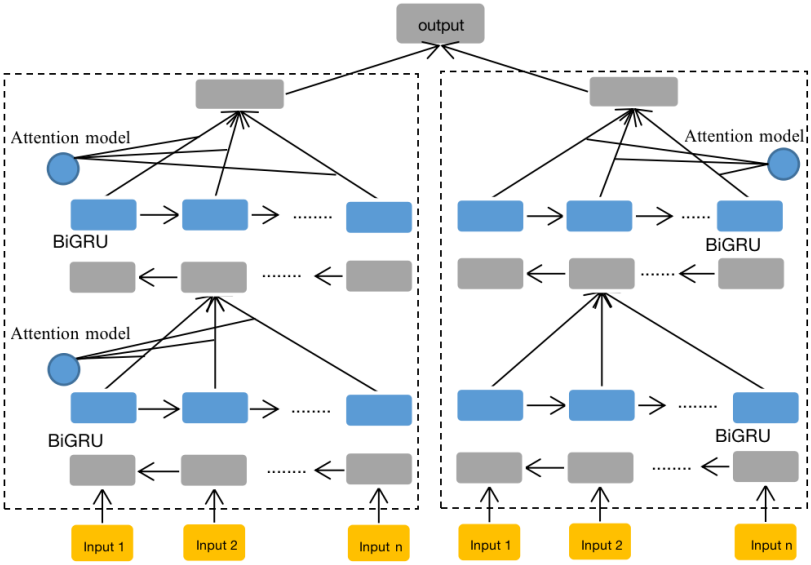

Figure 1. Fine-grained attention network architecture

In the figure 1, we can see that the model uses two sub-model channels, and different attentions are used for different channels. The data of hierarchical attention channel need to be modeled and calculated twice:

$$output\_word = f(word\_input) \tag{5}$$

$$\tilde{h}\_word = \sum_j a\_word_j \times ouput\_word_j \tag{6}$$

$$output\_sent = f(\tilde{h}\_word) \tag{7}$$

$$\tilde{h}\_sent = \sum_j a\_sent_j \times ouput\_sent_j \tag{8}$$

Output_word is the feature output of learning word level, $\tilde{h}\_word$ is the result of word level attention mechanism, output_sent is the feature output of sentence level, and $\tilde{h}\_sent$ is the result of sentence level attention mechanism. Then we need to fuse the different features of the two models together. Here we use average softmax function to fuse the models:

$$score_k = \sum_i h_{i,k}^T h_{i,k} \tag{9}$$

$$w_k = \frac{\exp(score_k)}{\sum_k \exp(score_k)} \tag{10}$$

$$output = w_1 \times h^1 + w_2 \times h^2 \tag{11}$$

h represents the original matrix of the model output; k represents the th-k record; i represents the th-i record; w represents the proportion of each model; output is the mixed output of the model.

4. Experiment

4.1 Experimental Parameters:

JingDong data of different commodity types, such as mouse, televisions, mobile phones and laptops, are used as Chinese data source. 22504 data are totally crawled, of which 11581 are positive class and 10923 are negative class. The data set is allocated to training set and data set according to the ratio of 3:2. In order to verify the robustness of the model, an English IBDM data set is introduced as a comparison. There are 50,000 reviews in IMDB, of which both positive and negative comments accounted for half. Table 1 shows the distribution of sentiment across different corpus size.

Table 1. allocation data.

	Total	*positive*	negative
JD_train_set	13502	6919	6583
JD_test_set	9002	4662	4340
IMDB_train_set	25000	12500	12500
IMDB_test_set	25000	12500	12500

Firstly, in the data preprocessing stage, Chinese record is segmented by jieba segmentation tools, while English data is segmented according to the special symbols. Then constructs word vector library by Skip-gram model. The model use Naive Bayes algorithm to filter noise record which calculated the statistical probability of feature words. We chose "好" ("good") and "差" ("bad") as reference word to calculate similar vocabulary respectively.

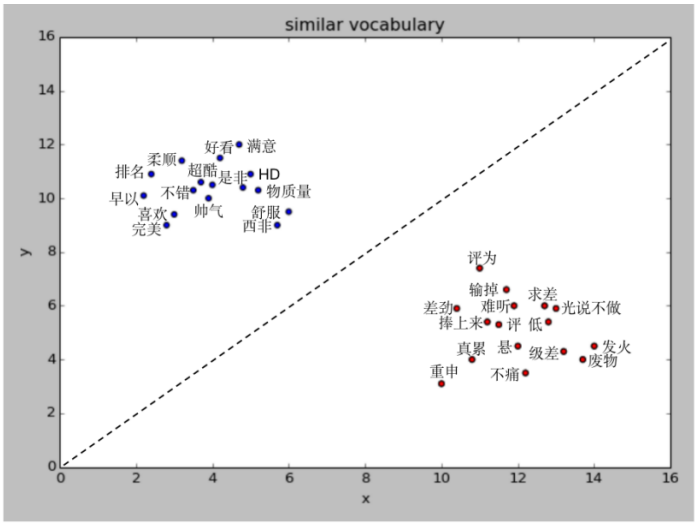

Figure 2. similar vocabulary

The similar word of "好" ("good") is 满意(Satisfaction), 柔顺 (suppleness), 好看 (good-looking), 排名(ranking), 是非 (right and wrong), 超酷(super cool), HD, 早以(early), 想不到 (unexpected), 不错(not bad), 物质量(material quality), 帅气(handsome), 喜欢(like), 完美(perfect), 舒服(comfortable), 西非(goods is very) .

The similar word of "差" ("bad") is 评为 (rating), 输掉 (losing), 评 (evaluating), 悬 (suspense), 差劲 (poor), 捧上来 (holding up), 废物 (trash), 难听 (offensive), 求差 (seeking difference), 光说不做 (just talking not doing), 真累 (really tired), 发火 (irritated), 级差 (over badly), 重申 (reiterating), 不痛 (painful)低(low). Mapping these words into two-dimensional space, the mapping results are shown in figure 2.

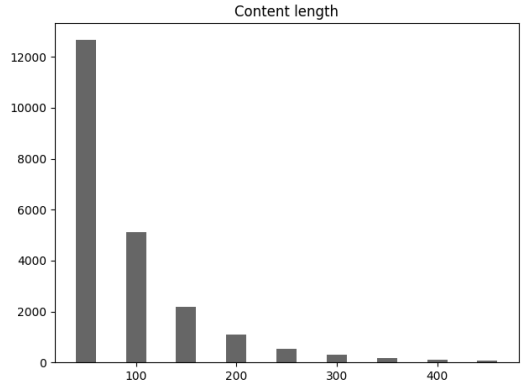

Figure 3. JD review length distribution, which abscissa is the length of the record and the ordinate is the number contained

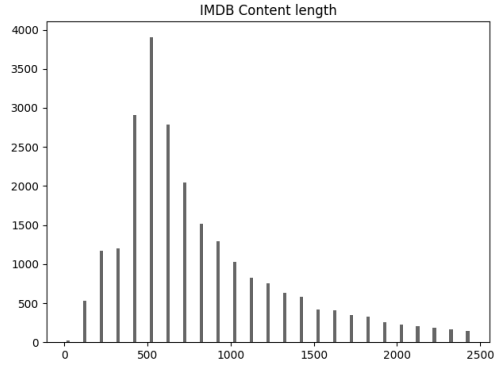

Figure 4. IMDB review length distribution, which abscissa is the length of the record and the ordinate is the number contained

For different data resources, the parameters of segment are different which depends on statistical information. As we can see from the figure 3 and figure 4, more than ninety percent of record in JD_corpus are within 300 length and 1800 for IMDB. Therefore, we setting the maximum sentence length of Chinese review and English review as 50 and 300, while fixed paragraph length as 6. The purpose of this preprocessing is to prevent the sparse vector matrix. The result before cutting is shown as figure 5 and the result after cutting is shown as figure 6.

In the process of experiment, some model parameters need to be determined manually. The influence of different parameter values on the model is compared by fixed parameter method. we chose parameter value as follows: echos setting 20,30,40; hidden layer size setting 90, 120, 150, 180; embedding dimension setting 10, 50, 100; dropout rate setting 0.2, 0.5, 0.8; learn_rate equals 0.0001,0.001, 0.01. The correct parameters involved in the model are obtained by experiment. The values are shown in the table 2.

Table 2. Model parameter values

parameter	Echos	*Batch_size*	Em_dimension	*Hidden_Size*	Dropout_rate	*Learn_rate*	*Threshold_pos*	*Threshold_neg*
Value	20	256	10	150	0.5	0.001	0.85	0.15

12	0.036255196	0.46694124	-0.86899686	1.8183851
13	0.5261831	0.1287386	-0.9123114	-1.409729
14	1.7211888	-2.2590485	-0.80781853	-1.3292335
15	1.8243291	-2.4729397	-3.6887567	-1.2189745
16	1.230795	-1.4442519	-1.7769331	-2.1958308
17	0.15173696	-0.18636885	-0.010597399	0.08636284
18	1.8251797	0.06359861	-2.447941	-0.27721488
19	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0

Figure 5. unprocessed data

12	0.036255196	0.46694124	-0.86899686	1.8183851
13	0.5261831	0.1287386	-0.9123114	-1.409729
14	1.7211888	-2.2590485	-0.80781853	-1.3292335
15	1.8243291	-2.4729397	-3.6887567	-1.2189745
16	1.230795	-1.4442519	-1.7769331	-2.1958308
17	0.15173696	-0.18636885	-0.010597399	0.08636284
18	1.8251797	0.06359861	-2.447941	-0.27721488
19	0.0	0.0	0.0	0.0

Figure 6. processed data

4.2 Results

The performance of FGANN-BiGRU model is compared with different deep learning models on different corpus. We selected traditional deep learn model and other effective deep learning model as the contrast models. To ensure that the results do not contain errors caused by accidental factors, each model run 10 times and the mean value is calculated as the final result.

The FGAtten-GRU model achieved accuracy surpassed other models' results on the same JD review sentiment dataset. Both the traditional deep learning model and the attention-GRU model plateaued at an accuracy just over 84%, with a maximum of 88.7%. We can find that our model has the highest performance which improves accuracy between 1.8% and 5.6% in JD corpus. The results of comparing different models are shown in table 3:

Table 3. Result in JD corpus.

	Accuracy	precision	F_score
RNN	84.9%	82.3%	86.1%
BiLSTM	86.7%	85.7%	87.5%
BiGRU	87.9%	85.7%	88.8%
Atten-BiGRU	88.7.%	87.7%	89.3%
FGAtten-BiGRU	**90.5%**	**92.8%**	**90.8%**

Our model's performance can also be compared to other methods in English corpus. Authors in [13] proposed a sentiment classification model with word attention based on weakly supervised CNN when they created the IMDB review corpus. Their highest reported accuracy was 87.29% from CNN model with hierarchical attention [13]. Authors in [14] experimented with multiple branches of deep learning models on the IMDB corpus. The best accuracy achieved in [14] was 89.5% from their combined CNN and LSTM model. As we can see from table 4, it is clear that our proposed method outperforms prior work.

Table 4. Result in IMDB corpus.

	Accuracy	loss
SVM	81.4%	0.410
Mixsupervised + BOW	88.9%	-
MBCNN-LSTM	89.5%	0.278
FGANN-BiGRU	**90.9%**	0.237

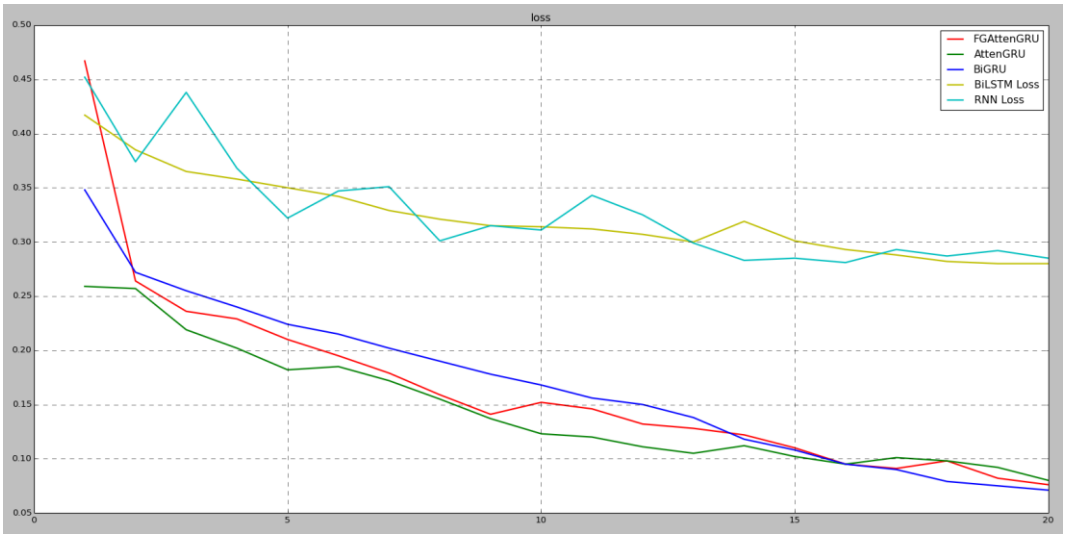

Figure. 7 loss

As can be seen from figure 7, the BiGRU, attention-GRU and FGANN-BiGRU model have better convergence effect. The model FGANN-BiGRU has no oscillation and the final loss value has dropped to a very low stable value, which has achieved better convergence effect.

5. Conclusion

In this paper, a feature extraction model based on fine-grained attention and bidirectional GRU network is proposed for sentiment classification of reviews. Through the integration of attention models at different granularity, our models are expected to learn more features of text reviews so as to improve the performance of the model. By using different corpus, we can find that Chinese data sets can reach 90.5% in accuracy, while English data sets can reach 90.9%; Compared with the conventional attention model, our fine-grained attention model improves the accuracy value by around 1.8%. Besides, we may conclude that our model is better than others as it effectively improves the accuracy of classification.

In the future, we will transform the two-category sentiment analysis task into multi-category task. At the same time, we will try to improve the method of segmenting Chinese words to make Chinese word segmentation more accurate, so as to improve the accuracy of Chinese data sets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61773037)

References

[1] Jieru Jia,Qiuqi Ruan,Gaoyun An,Yi Jin. 2017 Multiple metric learning with query adaptive weights and multi-task re-weighting for person re-identification[J]. Computer Vision and Image Understanding,160.

[2] Liu J, Zhang Y. 2017 Attention modeling for targeted sentiment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, p572-577.

[3] Asha S Manek, P Deepa Shenoy, et al. 2016 Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. World Wide Web, 20(2), 135-154.

[4] Kim Y 2014 Convolutional neural networks for sentence classification. EMNLP.

[5] Trofimovich J. 2016. Comparison of neural network architectures for sentiment analysis of russian tweets. Proceedings of the International Conference Dialogue 2016, RGGU.

[6] Bahdanau, Dzmitry,Cho, Kyunghyun,Bengio, Yoshua, 2014 Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473
[7] Zhaowei Qu, Yuan Wang, Xaioru Wang. A Hierarchical Attention Network Sentiment classification Algorithm Based on Transfer Learning[J]. Journal of Computer Applications, 2018, 38(11):3053-3056.
[8] Lijuan Zheng, Hongwei Wang, et al. 2015 Sentimental feature selection for sentiment analysis of Chinese online reviews[6]. Proceedings of the International Journal of Machine Learning and Cybernetics.
[9] Kalchbrenner N,Grefenstette E,Blunsom P 2014 A Convolutional Neural Network for Modelling Sentences. https://arxiv.org/abs/1404.2188.
[10] LI J,CAO Y,WANG Y,et al. 2017 Online learning algorithms for double-weighted least squares twin bounded support vector machines. Neural Processing Letters
[11] Liu J, Zhang Y. 2017 Attention modeling for targeted sentiment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, p572-577.
[12] Zhou X,Wan X, et al. 2016 Attention-based LSTM network for cross-lingual sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing p247-256.
[13] Lee G, Jeong J U, et al. 2018 Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowledge-Based Systems, S0950705118301710.
[14] Alec.Yenter, Abhishek.Verma, et al 2018 Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDb Review Sentiment Analysis, in 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON).