Building Corpus with Emoticons for Sentiment Analysis

Changliang Li[1(✉)], Yongguan Wang[2], Changsong Li[3], Ji Qi[1], and Pengyuan Liu[2]

[1] Kingsoft AI Laboratory, 33, Xiaoying West Road, Beijing 100085, China
{lichangliang,qijil}@kingsoft.com
[2] Beijing Language and Culture University,
15, Xueyuan Road, Beijing 100083, China
yongguanl992@163.com, liupengyuan@blcu.edu.cn
[3] Peking University, 5, Yiheyuan Road, Beijing 100871, China
lichangsong@pku.edu.cn

Abstract. Corpus is an essential resource for data driven natural language processing systems, especially for sentiment analysis. In recent years, people increasingly use emoticons on social media to express their emotions, attitudes or preferences. We believe that emoticons are a non-negligible feature of sentiment analysis tasks. However, few existing works focused on sentiment analysis with emoticons. And there are few related corpora with emoticons. In this paper, we create a large scale Chinese Emoticon Sentiment Corpus of Movies (CESCM). Different to other corpora, there are a wide variety of emoticons in this corpus. In addition, we did some baseline sentiment analysis work on CESCM. Experimental results show that emoticons do play an important role in sentiment analysis. Our goal is to make the corpus widely available, and we believe that it will offer great support to sentiment analysis research and emoticon research.

Keywords: Emoticon · Sentiment analysis · Corpus

1 Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinion and sentiments towards entities such as products, events and topics and so on [1].

The problem of sentiment analysis has been of great interest in the past decades because of its practical applicability. For example, consumers can seek advices about a product to make decisions in the consuming process. And vendors are paying more and more attention to online opinions about their products and services. Hence, sentiment analysis has attracted increasing attention from many research communities such as machine learning, data mining, and natural language processing.

Recently, people increasingly use emoticons on social media to express their feelings. The sentiment of a message is often affected by the emoticons that appear in the text. For example, given two movie reviews: "It's not Hollywood style :-)" and "It's

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 309–318, 2018.
https://doi.org/10.1007/978-3-319-99501-4_27

not Hollywood style :-(", though the text information is the same, but both reviews represent different sentiment due to different emoticons used.

For further research, we need large scale corpora with emoticons. However, most existing corpora are based on texts but ignore the emoticon information. This restricts the research for sentiment analysis and emoticon. So we established a large scale Chinese Emoticon Sentiment Corpus of Movies (CESCM). On CESCM, we have done a lot of experiments, and the experiment results show that the emoticon information help to analyze sentiment better (see the result in Sect. 4). This indicates that emoticons have great influence on the emotion of the whole text.

The remaining paper is structured as follows. Section 2 briefly discusses related work about sentiment dataset and emoticon. Section 3 describes the process of creating the corpus and gives the statistics of the corpus. Section 4 gives the experiment results based on CESCM. Section 5 presents a conclusion.

2 Related Work

Movie reviews are popular resource for sentiment analysis research. For example, Cornell Movie Review Data [2, 3] (MRD) is a commonly used sentiment analysis corpus that includes three datasets: sentiment polarity datasets, sentiment scale datasets, and subjectivity datasets. The Stanford Sentiment Treebank, referred as SST-1, is the first corpus with fully labeled parse trees. Based on the SST-1, SST-2 is created simply with neutral reviews removed from SST-1. And SST-2 is used for the binary classification task [4–6]. Stanford's Large Movie Review Dataset (LMRD) is used for binary sentiment classification and contains more substantial data compared with previous benchmark datasets [8]. Reviews from IMDB and YELP have also been used for some related research. And there is the Chinese Opinion Analysis Evaluation (COAE) [7], which contains product, movie, and finance reviews. Another is Chinese Sentiment Treebank [8, 9].

At present, some people have noticed the role of emoticon and carried out a series of work about emoticon. Someone consider emoticons as noisy labels in their sentiment classification models [10–12]. [13, 14] exploit emoticons in lexicon-based polarity classification. As an important research direction, emoticons are attracting more and more attention. So it is critical to building an corpus with emoticons to support sentiment analysis research.

Despite many corpora have been created for sentiment research, there still lacks large enough corpus with emoticons. For improving sentiment analysis research, we established a large scale Chinese Emoticon Sentiment Corpus of Movies (CESCM). In the next part, we will describe the process of creating the corpus and gives the statistics and analysis of the corpus.

3 Corpus Construction

In this section, we introduce the process of creating CESCM. First of all, we collect a lot of review data. Then, we build a common emoticon set and filter the movie review data based on the set. Next, we clean the data to get rid of redundant movie review data. Finally, we successfully constructed the corpus and analyzed it. We will describe the four key steps in detail.

3.1 Data Collection

Quality and scale are two most important aspects of a corpus. We collect the review data from famous Chinese movie review websites www.douban.com, which is the largest and best movie review site in China. On this website, there are a lot of short reviews about each movie. Each review is made up of one or several sentences. And each review has a star label. The number of star ranges from 1 to 5. People express their attitudes and preferences through movie reviews and star ratings. In addition to using text, there are many emoticons used in the reviews, such as emoticon ":)" or ": (". So we use crawler technology to collect the reviews and the corresponding labels from the website.

To study the usage of emoticon in text, we first constructed a common emoticon set. This emoticon set includes 510 emoticons that cover most of the popular emoticons on the web. We collected and collated these emoticons from the Internet by artificial methods and there are various styles of emoticons. For instance, emoticons in Western style have the eyes on the left, followed by nose and the mouth such as ":-)", and emoticons in Japanese style such as "(*_*)" (Also known as "Kaomojis"). In addition, we also selected some of the emoticons from the Unicode character set such as "😊". Some examples of the emoticon set were shown in Table 1.

We build CESCM based on the emoticon set. After we collected a great deal of movie reviews, we use these emoticons as necessary conditions to extract all reviews containing emoticon. In the end, we obtained a large number of movie review original resources with rich emoticon information.

3.2 Data Cleaning

The original data contains many false, dirty or unusable data. It can also be called "noise data". We utilize various strict rules for noise filtering to obtain high quality corpus data. There are several major rules below.

(1) We remove the reviews without star label or content.
(2) We remove the reviews unrelated to the movie. Such as website link, advertising and meaningless characters or gibberish.
(3) We limit reviews length from 3 to 50 words. Too long or too short will affect the quality of the corpus. Figure 1 gives the statistic of the movie reviews length after processing. We can see that the length of most reviews is between 5 and 20 words.

Note that we just list part of rules as example. There are more strict rules employed at this step.

Table 1. Review examples

ID	Emoticon	Description
1	= =	Awkward, indifferent, satiric
2	><	Annoyed, excited, uneasy, hesitant
3	=。=	Awkward, indifferent, tired
4	╮(╯▽╰)╭	Shrug
5	TAT	Sad, crying
6	XD	Laughing, big grin
7	T T	Sad, crying
8	:)	Smiley or happy face
9	O(∩_∩)O	Joyful, excited
10	╮(╯_╰)╭	Dissatisfied

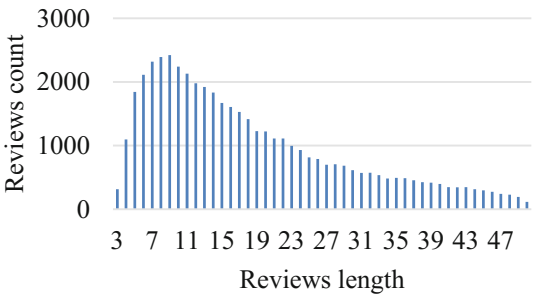

Fig. 1. The length distribution of reviews

3.3 Data Annotation

Table 2 shows some examples. To better understand the meaning, we translate the review into English.

In the CESCM, each review is labeled as stars, and the star number ranges from 1 to 5. These labels are given by people who have seen the movie and represent their evaluation of the movie. 1 means very negative; 2 means negative; 3 means neutral; 4 means positive; 5 means very positive.

Table 2. Review examples

Binary	class	Reviews
Negative	1	╮(╯▽╰)╭，没啥说的。 ╮(╯▽╰)╭, Have nothing to say.
	2	快睡着了，硬撑着看完的=。= Fast asleep, hard to finish watching =。=
	3	我觉得适合看文艺片= = It seems that I am suitable for art films = =.
Positive	4	这不是大侦探 这是谢耳朵！聪明是 21 世纪的性感:) This isn't a detective, it's Sheldon! Smart is 21st century sexy :)
	5	将近 4 小时的长片(^o^) 配乐超赞，越来越喜欢罗伯特德尼罗了！ It's nearly four hours long(^o^), The soundtrack is excellent, more and more like Robert DE Niro.

Besides fine-grained (five classes) analysis, this corpus can also be used for binary analysis. For this purpose, we marked each review with a new label based on its star label, just positive and negative. Reviews with 4 and 5 stars belong to positive, and others belong to negative.

3.4 Statistics and Analysis

After the process above, we obtain 47,250 high quality movie review data finally, and each movie review contains at least one emoticon. To our best knowledge, CESCM is the first large scale Chinese sentiment corpus with emoticons. We do a comparison with other popular sentiment analysis corpus, which are widely employed in sentiment analysis. The statistical summary of these corpora is presented in Table 3.

Compared to others, CESCM is larger and contains abundant information of emoticons. And each review contains at least one emoticon, and some reviews contain multiple different emoticons.

In order to better understand CESCM, Table 4 gives the statistic of different emoticons in each movie review. We can see that most reviews contain only one emoticon. There are 3374 reviews contain two emoticons, but only a few number of reviews contain 3 or more emoticons.

Table 3. The size of corpora

Corpora	Total size
MR [2]	10,662
SST-1 [4]	11,855
SST-2 [4]	9,613
Sentiment-Li [8]	13,550
CESCM	47,250

Table 4. Statistic of the emoticon count in each review

#Emoticon/Review	Reviews count
1	43683
2	3374
3	183
4	9
5	1

Figure 2 gives the distribution of emoticons. There are in total 175 different emoticons appear in CESCM. Figure 2 shows the proportion of top 15 emoticons and other emoticons. We can see that the emoticon "= =" is used most. And we found that people were more likely to use this kind of emoticons with rich meaning in real-world.

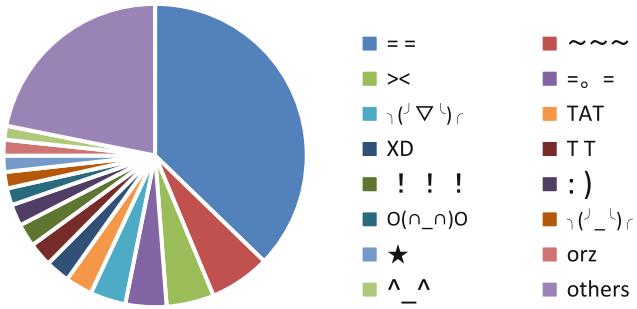

Fig. 2. Emoticons distribution in CESCM

Figure 3 gives an example of the distribution of emoticon sense. We chose a typical emoticon ":)" as illustration. From Fig. 3, we can see that ":)" is mostly employed as positive token. It should be noted that different emoticons may have different distribution of emoticon sense.

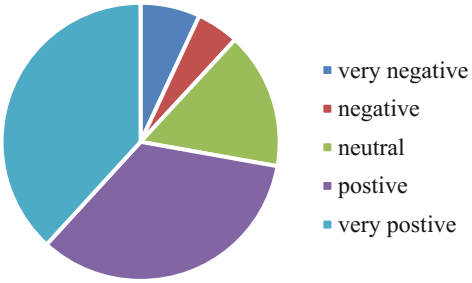

Fig. 3. Emoticon ":)" usage distribution

4 Experiment

In this section, we design experiment combination with our corpus. And we also introduce some baseline results on our corpus for other researchers to compare, as well as to understand the nature of the corpus. These approaches are widely used as baselines in sentiment analysis work, We report the experiment result on two tasks: fine-grained (5-class) and binary analysis (2-class).

4.1 Methods

Majority Method. It is a basic baseline method, which assigns the majority sentiment label in training set to each instance in the test set.

Feature Based Method
NB. We train a Naive Bayes classifier with TFIDF as text feature.
SVM. We train a SVM classifier with TFIDF as text feature.

Neural Network Method
Fast-text. Fast-text is a simple model for sentence representation and classification [15].
CNN. This approach utilizes convolutional neural networks (CNN) for sentiment classification just like work [16]. We employed max and kmax pooling respectively.
LSTM. LSTM based model is applied from the start to the end of a sentence. LSTM uses the last hidden vector as the sentence representation [17, 18]. LSTM_2 used 2-layer LSTM and Bi-LSTM represents bidirectional LSTM model.
Models with Emoticon Information. For baseline purpose, we just add the emoticon information on reviews simply as input fed to the models mentioned above.

4.2 Result

Table 5 gives experiment results for fine grained (5-class) and binary (2-class) sentiment predictions. The metric is accurate rate of predicted sentiment label.

Our experiment was divided into two groups, one group uses emoticons and the other does not.

Table 5. Experiment results of different approach

Information	Method	Fine-grained	Binary
Without emoticons	Majority	27.03%	50.88%
	NB	35.68%	67.09%
	SVM	34.98%	66.94%
	Fast-text	39.47%	77.86%
	CNN_max	40.99%	78.12%
	CNN_kmax	40.79%	78.16%
	LSTM	39.97%	78.62%
	LSTM_2	40.10%	78.29%
	Bi-LSTM	41.35%	78.31%
With emoticons	Majority	27.03%	50.88%
	NB	38.50%	70.20%
	SVM	37.30%	71.30%
	Fast-text	39.55%	79.09%
	CNN_max	42.39%	78.64%
	CNN_kmax	42.78%	78.62%
	LSTM	42.47%	79.01%
	LSTM_2	42.24%	80.04%
	Bi-LSTM	**43.60%**	**80.13%**

In both groups, majority performs worst, because it is just a simple statistic method without any semantic analysis. NB and SVM perform similar as a representative of machine learning methods. However, the performance of NB and SVM is poor compared with the method based on neural network. In the method of neural network, LSTM performs generally better than CNN and Fast-text.

From comparison between two groups, we can see that the methods in second group outperforms the corresponding method in group one. For CNN-based models, it usually increases by more than 1.4% in fine-grained and 0.4% in binary. For LSTM-based models, it usually increases by more than 2.1% in fine-grained and 0.3% in binary. This proves that emoticons play an important role in emotional analysis.

From the result, we can see that combination with the emoticon information boosts the performance of some simple methods. This not only proves the importance of emoticons, but also the potential of our corpus in both sentiment analysis research and emoticon analysis.

5 Conclusion

In this paper, we introduce a high quality and large sentiment corpus of Chinese movie reviews (CESCM). The corpus consists of 47,250 reviews with emoticon information. By explaining the data creation process and providing the results of the baseline algorithms, we hope to help researchers to better understand the nature of CESCM. We

believe that the corpus described in this paper will offer great help to researchers for future research on emoticons and sentiment analysis.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2007)
2. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124 (2005)
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
4. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
5. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through Recursive matrix-vector spaces. In: Empirical Methods in Natural Language Processing, pp. 1201–1211 (2012)
6. Socher, R., Pennington, J., Huang, E., Ng, A.Y., Manning, C.D.: Semi-Supervised recursive autoencoders for predicting sentiment distributions. In: Empirical Methods in Natural Language Processing, pp. 151–161 (2011)
7. 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, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
8. Li, C., Xu, B., Wu, G., He, S., Tian, G., Hao, H.: Recursive deep learning for sentiment analysis over social data. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, pp. 180–185. IEEE Computer Society (2014)
9. Li, C., Xu, B., Wu, G., He, S., Tian, G., Zhou, Y.: Parallel recursive deep model for sentiment analysis. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015, Part II. LNCS (LNAI), vol. 9078, pp. 15–26. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18032-8_2
10. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical report (2009)
11. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: 7th Conference on International Language Resources and Evaluation (LREC 2010), pp. 1320–1326. European Language Resources Association (2010)
12. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: AAAI Conference on Artificial Intelligence (2012)
13. Hogenboom, A., Bal, D., Frasincar, F., et al.: Exploiting emoticons in sentiment analysis. In: ACM Symposium on Applied Computing, pp. 703–710. ACM (2013)
14. Hogenboom, A., Bal, D., Frasincar, F., et al.: Exploiting emoticons in polarity classification of text. J. Web Eng. **14**(1–2), 22–40 (2015)
15. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification (2016). arXiv preprint: arXiv:1607.01759

16. Kim, Y.: Convolutional neural networks for sentence classification. In: Empirical Methods in Natural Language Processing, pp. 1746–1751 (2014)
17. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: Interspeech, vol. 31, pp. 601–608 (2012)
18. Wang, Y., Feng, S., Wang, D., Zhang, Y., Yu, G.: Context-aware Chinese microblog sentiment classification with bidirectional LSTM. In: Li, F., Shim, K., Zheng, K., Liu, G. (eds.) APWeb 2016, Part I. LNCS, vol. 9931, pp. 594–606. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45814-4_48