Sentiment Analysis for Tang Poetry Based on Imagery Aided and Classifier Fusion

Yabo Shen[1], Yong Ma[1], Chunguo Li[2], Shidang Li[1], Mingliang Gu[1(✉)], Chaojin Zhang[1], Yun Jin[1,3], and Yingli Shen[1]

[1] School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou, China
shenyabohpu@163.com, mlgu@jsnu.edu.cn
[2] School of Information Science and Engineering, Southeast University, Nanjing, China
[3] Kewen College, Jiangsu Normal University, Xuzhou, Jiangsu, China

Abstract. This paper aims to do sentiment analysis for Tang poetry from the perspective of text mining. Most previous works just focus on the literariness of Chinese poetry or establish language models statistically, which ignore the features of sentiment and specific applications. We propose a sentiment analysis system for Tang poetry based on imagery aided and classifier fusion. Especially, we extract sentimental imageries at two levels: character and word, and bring them into sentiment analysis. In addition, classifier fusion is adopted in this paper to improve classification performance. Experiments show the effectiveness of our model and our method is superior to the traditional method.

Keywords: Sentiment analysis · Tang poetry · Imagery aided · Classifier fusion

1 Introduction

As a kind of classical Chinese poetry, Tang poetry is precious spirit wealth of human being [1]. Nowadays, with the rise of artificial intelligence, it becomes a hot trend to process and analyze various data through machine learning or natural language processing (NLP) [2]. However, for classical Chinese poetry, most related works just focus on its literariness or establish language models locally while ignoring the global features of statistics and the applications in NLP.

In order to apply NLP to classical Chinese poetry, Fang [3] addressed the issue of machine-aided analysis and understanding of classical Chinese poetry and proposed a computational framework. Nevertheless, it suffered from high model complexity and focused more on the imagery itself. He [4] used general methods of machine learning, and established a SVM-based poetry style classification system while there was too little detailed study.

Besides, Liu [5] studied colors in Tang poetry and the social networks of the Tang poets. Wang [6] proposed a novel two-stage poetry generating method and Zhang [7] introduced a model for Chinese poetry generation based on recurrent neural networks.

S. Han et al. (Eds.): AICON 2019, LNICST 287, pp. 283–290, 2019.
https://doi.org/10.1007/978-3-030-22971-9_24

Peng [8] studied the multi-grained Chinese text representation for Chinese sentiment analysis. Actually, all these studies give us great inspiration and help.

In this paper, we select fine-grained Tang poetry texts, which have been manually marked, and then perform sentiment analysis task. In addition to the general framework of sentiment analysis, we also make statistical analysis of the Complete Tang Poems (CTP) and extract sentimental imageries at character and word levels to help us implement the task. Furthermore, classifier fusion is used in this paper to optimize the performance. The contributions of our work can be summarized as follows:

(1) This paper makes statistical analysis of the CTP (about 43000 poems) from the character and word levels to extract the sentimental imageries and brings them into classification model.
(2) The idea of sentiment analysis for short text is transferred to the Tang poetry, and classifier fusion is adopted to obtain better experimental results.

The remainder of the paper is organized as follows. Section 2 presents our model and corresponding analysis in detail. Section 3 describes the experimental settings and results. In Sect. 4, the conclusion is drawn.

2 Our Proposed Method

In this section, we present the details of our sentiment analysis system for Tang poetry. An overall framework is shown in Fig. 1.

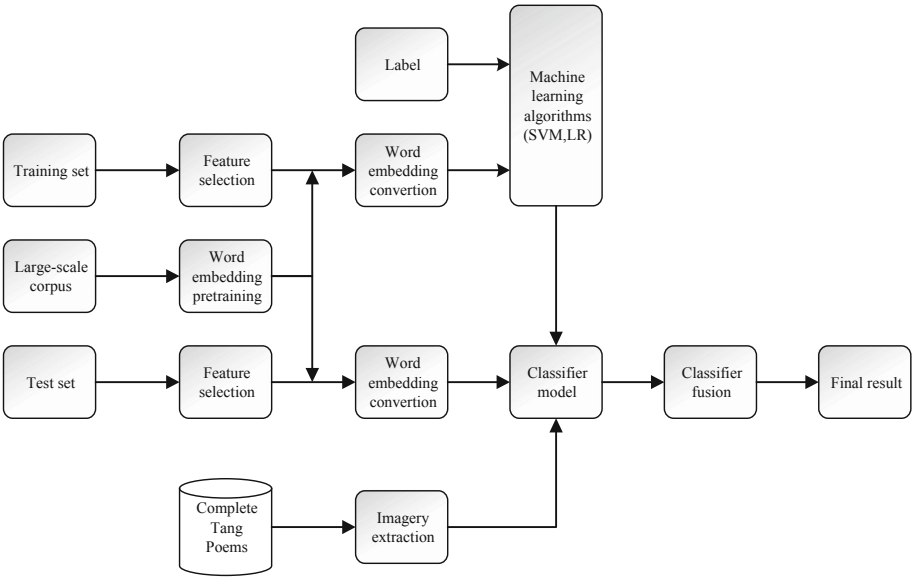

Fig. 1. The framework of sentiment analysis system for Tang poetry

2.1 Sentiment Analysis for Tang Poetry

Text sentiment analysis refers to the process of detecting, analyzing and mining subjective texts including views, preferences and emotions expressed by users [9]. Actually, it is a specific functional implementation of text classification.

As shown in Fig. 1, the sentiment analysis system for Tang poetry consists of training and testing part. For the labeled training set, we extract the features by information gain (IG) and mutual information (MI).

Information gain is used to calculate the amount of information contributed to the text classification by the presence or absence of features, and the formula is as follows:

$$Gain(w_i) = Entropy(S) - Entropy(S_i) \tag{1}$$

Entropy(S) indicates the information entropy when the feature w_i does not appear, and $Entropy(S_i)$ indicates the information entropy after the appearance of the feature w_i.

Mutual information is used in text classification to measure how much a feature is dependent on a category, as follows:

$$MI(w_i, c_j) = \log \frac{p(w_i, c_j)}{p(w_i)p(c_j)} \tag{2}$$

$p(w_i, c_j)$ indicates the probability that a document containing feature w_i belongs to class c_j, and $p(w_i)$ indicates the probability that feature w_i will appear in the document, and similarly $p(c_j)$ indicates the probability that a document belongs to class c_j.

Then according to the pretrained word embeddings, we convert the features extracted to word embeddings of corresponding dimensions. Next, what we do is establishing the classifier model using machine learning algorithms (i.e. SVM, LR).

In the same way, for the test set, feature extraction and word embedding convertion are implemented, and when we give a testing sample, the classifier model will return a label as "1" or "0" (positive = 1, negative = 0).

2.2 Word Segmentation for Tang Poetry

Word segmentation is the basis of Chinese Natural Language Processing and has considerable importance [10]. If word segmentation is not good, the following result is probably not good either.

For example, if we make word segmentation for the poem "烽火连三月，家书抵万金 (feng huo lian san yue, jia shu di wan jin)". The correct segmentation should be "烽火 连 三月, 家 书 抵 万金", which shows the cruelty and persistence of the war, as well as the eagerness to get family news. But if divided into "烽火 连 三月, 家 书 抵 万 金", it will make people very confused about the meaning of its expression.

In this paper, we use THULAC [11] to make word segmentation.

2.3 Imagery Aided Classification

Imagery in poetry refers to a meaningful image, that is, an artistic image created through the unique emotional activities of the creative subject. And if we know the imagery well, we can easily judge the sentiment of the poetry.

As shown in Fig. 2, text preprocessing is the first step because the text contains ordinal numbers, titles, authors and contents. So it is segmented and preliminarily screened out. At the same time, special symbols should be eliminated. We only keep commas and periods, and the rest of symbols are regarded as invalid symbols. Then, for the preprocessed text, we make statistical analysis at two levels. For character level, we count character frequency and find some characters that can be imageries. And for word level, we make word segmentation and statistical analysis. By these steps, we can get the sentimental imageries we need.

Thus we can use the sentimental imageries to construct a classifier for Tang poetry according to the imageries included in the poems.

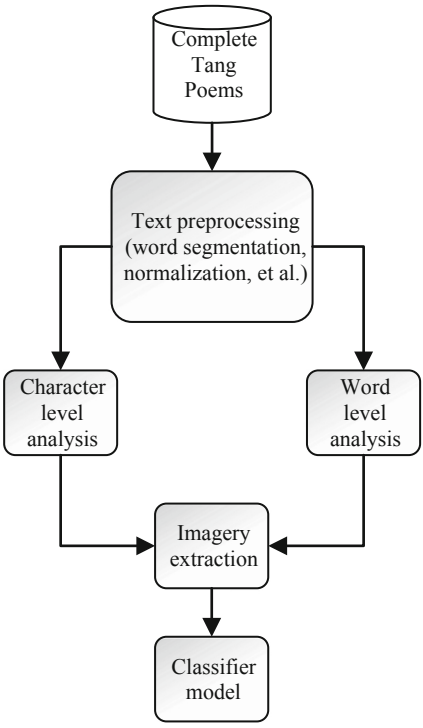

Fig. 2. The detailed process of imagery aided classification

2.4 Classifier Fusion

In order to improve the performance of the proposed method, the classifier fusion method is adopted [12], and the detail is shown in Fig. 3.

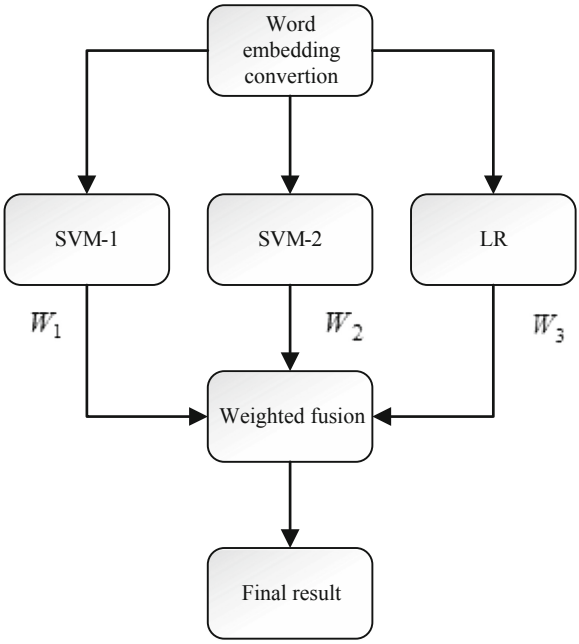

Fig. 3. The detailed process of classifier fusion module

As shown in Fig. 3, three independent classifiers are used. The formula is as follows:

$$P = \sum_{i=1}^{3} W_i * P_i \tag{3}$$

The weight of each classifier W_i(i = 1, 2, 3) is determined by grid search algorithm. Then the weights are multiplied by corresponding probability P_i and sum to obtain the final probability p.

3 Experiments Analysis

3.1 Experiment Setting

As shown in Table 1, the experimental data in this paper are the untagged Complete Tang Poems (a text of 42974 poems) and the labeled 5598 lines of Tang poems. We label 5598 lines of poems manually, among which 3,403 are negative and 2,195 are positive.

The experiments in this paper are all carried out in python environment of Linux system.

Table 1. Experimental data and environment

	Data
1	Complete Tang Poems (42974 poems)
2	labeled Tang poems (5598 lines)

In this paper, SVMs and LR (Logistic Regression) are used for the classifier fusion (Table 2).

Table 2. Classifier setting

Classifier 1	Classifier 2	Classifier 3
SVM-1 (kernel = 'rbf', degree = 2, gamma = 1, coef0 = 10)	SVM-2 (kernel = 'poly', degree = 2, gamma = 10, coef0 = 10)	LR (penalty = '12')

In this paper, the accuracy (Acc) is used to evaluate the performance of the classifiers. It represents the proportion of the total number of correctly predicted samples in the test samples to the total number of actual samples. In general, the higher the accuracy is, the better the classifier's performance will be. The formula is as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

TP means that the predicted results are consistent with the actual results and are positive samples. TN means that the predicted results and actual samples both are negative. FP means that the predicted results are positive but the actual samples are negative. Similarly, FN means the opposite of FP.

3.2 Result Analysis

As can be seen from Table 3, through the statistical analysis of Tang poetry at the character and word level, we find some sentimental imageries that can be used to identify the sentiment polarity (not all listed).

Table 3. Sentimental imageries

	Positive Imageries		Negative Imageries	
	Chinese	*English*	*Chinese*	*English*
Character level	松	pine	柳	willow
	竹	bamboo	猿	ape
	梅	plum	秋	autumn
	兰	orchid	坟	tomb
Word level	春风	Spring breeze	白发	white hair
	朝阳	the rising sun	白骨	bones of the dead
	高山	high mountain	夕阳	the setting sun
	青天	blue sky	孤城	lonely town
	青云	high official position	落花	falling flowers
	碧水	blue water	琵琶	Chinese lute

Table 4. Tang poetry's sentiment analysis results (dimension = 100)

Method		Acc
No classifier fusion	Imagery-aided classifier	52.4%
	SVM-1	58.6%
	SVM-2	54.8%
	LR	59.0%
With classifier fusion		**67.2%**

As can be seen in Table 4, the accuracy of sentiment classification with classifier fusion is obviously higher than that without classifier fusion.

Table 5. The results when word embedding dimensions change

Dimension	Acc
100	67.2%
200	**67.9%**
300	67.5%
400	66.9%

Besides, we also explore the influence of the word embedding dimension. From Table 5, it doesn't make much difference due to the characteristic of short text.

4 Conclusion

This paper proposes a sentiment analysis system for Tang poetry based on imagery aided and classifier fusion. The results show the effectiveness of the system, which is of certain significance to the research on the sentiment analysis for Tang poetry and other literary short texts. There are still some problems in this paper, such as the low accuracy. All in all, the method in this paper is worth further research in the future.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Grant No. 61708061, 6167114, and 61673108), Xuzhou Science and technology project (KC18015), Industry-University-Research collaboration project of Jiangsu Province (BY2018077), Research Fund for the Doctoral Program of New Teachers of Jiangsu Normal University (Grant No. 17XLR029), Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 17KJB510016, 17KJB510018, and 18KJB510013), and Jiangsu Normal University School Funding Project (2018YXJ611).

References

1. Hou, Y., Frank, A.: Analyzing sentiment in classical chinese poetry. In: Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 15–24 (2015)
2. Zheng, Y.: Affective computing applied in chinese classical poetry study. E-sci. Technol. Appl. **3**(4), 59–66 (2012)
3. Fang, A.C., Lo, F., Chinn, C.K.: Adapting NLP and corpus analysis techniques to structured imagery analysis in classical chinese poetry. In: Workshop Adaptation of Language Resources and Technology to New Domains, pp. 27–34 (2009)
4. He, Z., Liang, W., Li, L., et al.: SVM-based classification method for poetry style. In: the Proceedings of ICMLC06, pp. 1588–1591 (2007)
5. Liu, C., Wang, H., Cheng, W., et al.: Color aesthetics and social networks in complete tang poems: explorations and discoveries. Comput. Sci. (2015)
6. Wang, Z., He, W., Wu, H., et al.: Chinese poetry generation with planning based neural network. In: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1051–1060 (2016)
7. Zhang, X., Lapata, M.: Chinese poetry generation with recurrent neural networks. In: Proceedings of the 2014 Conference of Empirical Methods in Natural Language Processing (EMNLP), pp. 670–680 (2014)
8. Peng, H., Ma, Y., Li, Y., Cambria, E.: Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl.-Based Syst. **148**, 167–176 (2018)
9. Poria, S., Cambria, E., Bajpai, R., et al.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
10. Li, Y., Pan, Q., Yang, T., et al.: Learning word representations for sentiment analysis. Cogn. Comput. **9**, 843–851 (2017)
11. Sun, M., Chen, X., Zhang, K., Guo, Z., Liu, Z.: THULAC: an efficient lexical analyzer for Chinese (2016)
12. Zhao, D., Shen, Y., Shen, Y., et al.: Short text sentiment analysis based on windowed word vector. In: The 7th International Conference on Communication Signal Processing and Systems, p. 346 (2018)