A Joint Model for Chinese Microblog Sentiment Analysis

Yuhui Cao, Zhao Chen, Ruifeng Xu*, Tao Chen and Lin Gui
Shenzhen Engineering Laboratory of Performance Robots at Digital Stage,
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
caoyuhuiszu@gmail.com xuruifeng@hitsz.edu.cn

Abstract

Topic-based sentiment analysis for Chinese microblog aims to identify the user attitude on specified topics. In this paper, we propose a joint model by incorporating Support Vector Machines (SVM) and deep neural network to improve the performance of sentiment analysis. Firstly, a SVM Classifier is constructed using N-gram, N-POS and sentiment lexicons features. Meanwhile, a convolutional neural network is applied to learn paragraph representation features as the input of another SVM classifier. The classification results outputted by these two classifiers are merged as the final classification results. The evaluations on the SIGHAN-8 Topic-based Chinese microblog sentiment analysis task show that our proposed approach achieves the second rank on micro average F1 and the fourth rank on macro average F1 among a total of 13 submitted systems.

1 Introduction

With the development of the Internet, microblog has become a popular user-generated content platform where users share the newest events or their personal feelings with each other. Topic-based microblogs are the most common interactive way for users to share their opinions towards a specified topic. To identify the opinions of users, sentiment analysis techniques are investigated to classify texts into different categorizations according to their sentiment polarities.

Most existing sentiment classification techniques are based on machine learning algorithms, such as Support Vector Machine, Naïve Bayes and Maximum Entropy. The machine learning based approach uses feature vectors as the input of classification to predict the classification results. Thus, feature engineering, a method for extracting effective features from texts, plays an important role. Some commonly used features in sentiment classification are unigram, bigram and sentiment words. However, these features cannot work well for cross-domain sentiment classification because of the lack of domain knowledge.

Danushka Bollegala et al. (2011) used multiple sources to construct a sentiment sensitive thesaurus to overcome the lack of domain knowledge. New sentiment words expansion is another kind of approach to improve the performance of sentiment analysis. Strfano Baccianella et al. (2010) constructed SentiWordNet by extending WordNet with sentiment information. It is now widely used in sentiment classification for English. As for Chinese sentiment analysis, Minlie Huang et al. (2014) proposed a new word detection method by mining the frequent sentiment word patterns. This method may discover new sentiment words from a large scale of unlabeled texts.

With the rapid development of pre-trained word embedding and deep neural networks, a new way to represent texts and features is devloped. Mikolov et al. (2013) showed that word embedding represents words with meaningful syntactic and semantic information effectively. Recursive neural network proposed by Socher et al. (2011a; 2011b; 2013) is shown efficient to construct sentence representations based on the word embedding. Convolutional neural networks (CNN), another deep learn model which achieved success in image recognition field, was applied to nature language processing with word embed-

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 61–67,
Beijing, China, July 30-31, 2015.

dings. Yoon Kim (2014) used CNN with pre-trained word embedding to achieve state-of-the-art performances on some sentence classification tasks, including sentiment classification. Siwei Lai et al. (2015) incorporated global information in a recurrent convolutional neural network. It obtained further improvements comparing to other deep learning models.

In this paper, we propose a joint model which incorporates traditional machine learning based method (SVM) and deep learning model. Two different classifiers are developed. One is a word feature based SVM classifier which uses word unigram, bigram and sentiment words as features. Another one is a CNN-based SVM classifier which takes paragraph representations features learned by CNN as input features. The classification results of these two classifiers are integrated to generate the final classification results. The evaluations on the SIGHAN-8 Topic-based Chinese microblog sentiment analysis task show that our proposed approach achieves the second rank on micro average F1 and the fourth rank on macro average F1 among a total of 13 submitted systems. Furthermore, the joint classifier strategy brings further performance improvement on individual classifiers.

The rest of this paper is organized as follows. Section 2 presents the design and implementation of our proposed joint model. Section 3 gives the evaluation results and discussions. Finally, Section 4 gives the conclusion and future research directions.

2 Our Approach

The SIGHAN8 topic-based Chinese polarity classification task aims to is to classify Chinese microblog into three topic-related sentiment classes, namely neutral, positive and negative. This task may be generally regarded as a three-category classification problem. The SVM classifier which has been shown effective to document classification is adopted as the core classifier. Here, two different feature representation models, namely word-based vector space model and CNN-based composition representation, are adopted to generate the classification features for two classifiers, respectively. The classification outputs of two classifiers are integrated to generate the final output.

2.1 Data preprocessing

Chinese microblog text is obviously different from formal text. Many microblogs have noises, including nickname, hashtag, repost or reply symbols, and URL. Therefore, before the feature representation and extraction, preprocessing is performed to filter out noise text in the microblogs. Meanwhile, the advertising text and topic-irrelevant microblog are identified as neutral text. Especially, this task is designed to identify the topic-relevant sentiments. Therefore, the information coming from the reply, repost and sharing parts should be filtered out to avoid their influences to the sentiment analysis of the microblog author. Generally speaking, such filtering is based on rules. The table 1 shows the example data preprocessing rules with illustrations.

Table 2 shows the rules for identifying the advertisement and topic-irrelevant microblogs. The identified microblogs are labeled as neutral for topic-based sentiment classification.

2.2 Word feature based classifier

The word feature based classifier is designed based on the vector model. Firstly, the new sentiment words from unlabeled sentences data are recognized to expand the sentiment lexicon. The classification features are extracted from the labeled training data and sentiment lexicon resources. In order to alleviate the influences of unbalanced training data, SMOTE, which is an oversampling algorithm, is applied to training data before classifier training. Finally, a SVM classifier is trained on the balanced data. The framework of word feature based classifier is shown in Figure 1.

2.2.1 Feature selection

Unigram, Bigram, Uni-Part-of-Speech and Bi-Part-of-Speech features are selected as the basic features. CHI-test based feature selection is applied to obtain the top 20000 features. To improve the performance of sentiment classification, additional features based on lexicons including sentiment word lexicons, negation word lexicons, and adverb word lexicons, are incorporated.

Rules	Raw Text	Processed Text
Sharing news with personal comments	好看？吗？//【Galaxy S6：三星证明自己能做出好看的手机】http://t.cn/RwHRsIb(分享自 @ 今日头条)	好看？吗？
Removing HashTag	# 三星 Galaxy S6# 三星 GALAXY S6，挺中意 [酷][酷] [位置] 芒砀路	三星 GALAXY S6，挺中意 [酷][酷]
Removing URL	699 欧元起传三星 Galaxy S6/S6 Edge 售价获证实（分享自 @ 新浪科技）http://t.cn/RwTo3on	699 欧元起传三星 Galaxy S6/S6 Edge 售价获证实（分享自 @ 新浪科技）
Removing nickname	玻璃取代塑料，更美 Galaxy S6 的 5 大妥协 http://t.cn/RwHY6Az 罗永浩我去小米和三星这是要闹哪样,,, 老罗。。不能忍啊,,,,, @ 锤子科技营销帐号 @ 罗永浩	http://t.cn/RwHY6Az 罗永浩我去小米和三星这是要闹哪样,,, 老罗。。不能忍啊,,,,,
Removing information sources	【视频：三星 S6 对比苹果 iPhone6 MWC2015 @youtube 科技 】http://t.cn/RwHQzJ8（来自于优酷安卓客户端）	【视频：三星 S6 对比苹果 iPhone6 MWC2015 @youtube 科技 】http://t.cn/RwHQzJ8

Table 1: Data preprocessing rules with illustrations.

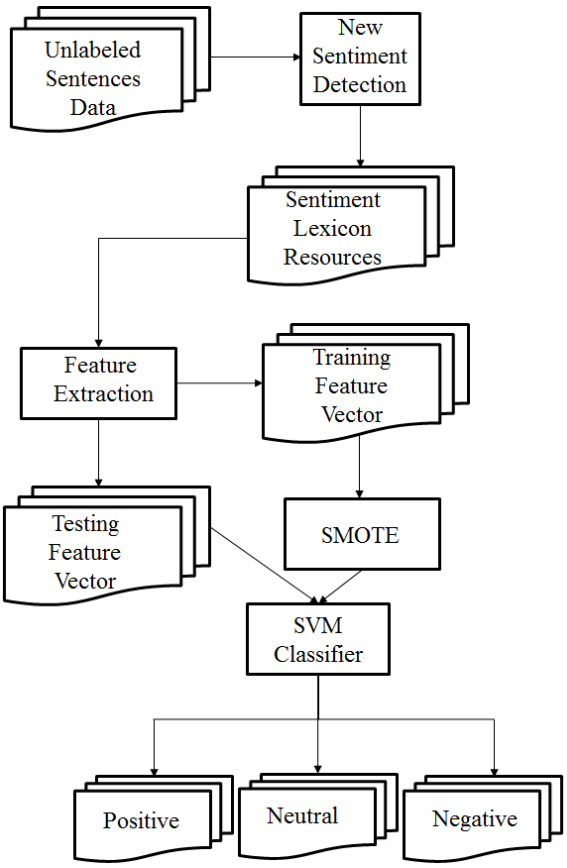

Figure 1: Framework of word feature based classifier

Rules	Type
Including many different topic (“#...#”) tag.	Advertisement
Including many words like “微商”, “商机”, “想赚钱”, “面膜”.	Advertisement
No actual content	Topic-irrelevant

Table 2: Microblog text matching rules.

By analyzing the expressions of the microblog text in training data, some special expression features in microblog text are identified. For example, the continuous punctuations are always used to express a strong feeling and thus, the microblog with continuous punctuations tends to be subjective. Another adopted feature for microblog text is the use of emoticons.

2.2.2 Sentiment lexicon expansion

In microblogs, abundant new or informal sentiment words are widely used. Normally, these new sentiment words are short but meaningful for expressing a strong feeling. These new sentiment words play an important role in Chinese microblog sentiment classification. Therefore, sentiment word identification is performed to recognize new sentiment words as the supplement of sentiment lexicon.

Twenty million microblog text collected from Sina Weibo Platform are used in new sentiment word detection. Considering that new words normally cannot be correctly segmented by the existing segmentor, identifying new words from preliminary segmentation results together with their POS tags is a feasible method. Here, potential components for new words are limited to the segmentation tokens shorter than three. Using word frequency, mutual information and context entropy as the evaluation indicators for words, the most possible new word candidates are obtained. With the help of word embedding construction model, each word in the corpus can be represented as a low dimension vector together with its context information. Hence, the distances between the new words and the existed sentiment words corresponding to difference sentiment polarity are estimated. The new words are then classified into one of the three polarity classes by following voting mechanism.

2.2.3 Classification

Two steps are performed to determine the topic-relevant sentiment for input microblogs. The first step is to distinguish topic relevant messages from topic irrelevant messages. Sentiment classification is then applied to topic relevant messages in the second step.

Topic relevant words generated by clustering analysis are employed as distinguishable features to filter out topic irrelevant microblogs because normally the topic irrelevant microblogs have few intersections with topic relevant words. Some advertisement posts consisting of several hot topic hash tags are also filtered out by considering the number of hash tag types in the microblog.

The provided labeled dataset is used to train the SVM classifier with linear kernel. A new challenge is that the provided training set is imbalanced. There are about 3973 neutral microblogs, while the numbers of positive and negative microblogs are 394 and 538, respectively. In order to reduce the influences of imbalanced training dataset, the SMOTE algorithm (Chawla et al., 2002) is applied to oversampling the samples on minority class. Oversampling ratio is set to 10 and 7.4 for positive class and negative class, respectively. In this way, the training dataset becomes balanced.

2.3 CNN-based SVM classifier

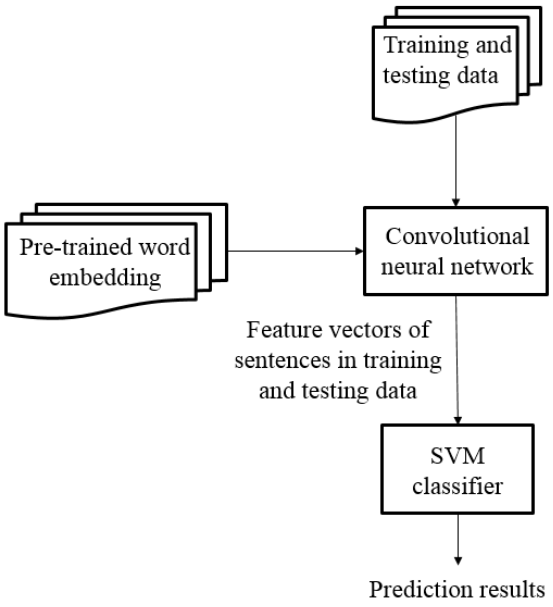

Figure 2: CNN and SVM joint classifier.

Another classifier is CNN-based SVM classifier. The classifier framework is shown in Figure 2. Firstly, continuous bog of word (CBOW) model (Mikolov et al., 2013) is used to learn word embeddings from Chinese microblog text. A deep convolutional neural networks (CNN) model is applied to learn distributed paragraph representation features for Chinese microblog training and testing data. Finally, the distributed paragraph representation features are used in SVM classifier to learn the probability distribution over sentiment labels.

2.3.1 Word embedding construction

Word embedding, wherein words are projected from a sparse, 1-of-V encoding (here V is the vocabulary size) onto a lower dimensional vector space via a hidden layer, are essentially feature extractors that encode semantic features of words in their dimensions. Mikolov et al. (2013) introduced CBOW model to learn vector representations which captures a large number of syntactic and semantic word relationships from unstructured text data. The main idea of this model is to find word representations which use the surrounding words in a sentence or a document to predict current word.

In this study, we train the CBOW model by using 16GB Chinese microblog text. Finally, we obtain 200-dimension word embeddings for Chinese microblog text.

2.3.2 CNN-based SVM classifier

In the CNN-based SVM classifier, the input is a matrix which is composed of the word embeddings of microblogs. There are windows with the lengths of three, four and five words, respectively. A convolution operation involves three filters which are applied to these windows to produce new features. After convolution operation, a max-over-time pooling operation is applied over these features. The maximum value is taken as the feature corresponding to this particular filter. The idea is to capture the most important feature which has the largest value. Since one feature is extracted from one filter, the model uses multiple filters (with varying window sizes) to obtain multiple features. These features constitute the distributed paragraph feature representation. In the last step, a SVM classifier is applied on these distributed paragraph representation features to obtain the probability distributions over labels (positive, negative, and neutral).

2.4 Outputs Merging

Classifier 1	Classifier 2	Final result
positive	neutral	neutral
negative	neutral	neutral
neutral	positive	neutral
neutral	negative	neutral
positive	negative	negative
negative	positive	positive

Table 3: Merging rules for two classifiers.

A set of merging rules is designed to incorporate the individual classification results of the two classifiers for generating the final result. If the two classification outputs are the same, naturally, the final output is the same. If the two classification outputs are different, the final result is determined from the merge rules shown in Table 3. Simply speaking, if any of two classifiers output neutral category, the final output is neutral. If two classifiers outputs positive and negative, respectively, the final output is the result of CNN-based classifier. Such a classification outputs merging strategy is based on the statistical analysis on the individual classifier performances on training dataset.

3 Experimental results and analysis

3.1 Data set

In the SIGHAN-8 Chinese sentiment analysis bakeoff dataset, 4905 topic-based Chinese microblog are provided as training data which consists of 394 positive, 538 negative and 3973 neutral microblogs corresponding to 5 topics, namely “央行降息”, “油价”, “日本马桶”, “三星 S6” and “雾霾”. In the testing data, there are 19,469 microblogs corresponding to 20 topic, such as “12306 验证码”, “中国政府也门撤侨”, “何以笙箫默”, “刘翔退役”.

3.2 Metrics

Precision, recall and F1-value are used as the evaluation metrics, as shown below:

$$Precision = \frac{SystemCorrect}{SystemOutput} \quad (1)$$

$$Recall = \frac{SystemCorrect}{HumanLabeled} \quad (2)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

Where $System.Output$ refers to the total number of the submitted results, $System.Correct$ refers to the number of correctly classified results in the submitted results, $Human.Labeled$ refers to the total number of manually labeled results in the Gold Standard.

The evaluation metrics corresponding to positive, negative and overall are estimated, respectively. The corresponding micro-average and macro-average performances are then estimated. The micro-average estimates the average performance of the three evaluation metrics over the entire dataset. The macro-average estimates the average performances of the evaluation metrics on positive, negative and neutral, respectively.

3.3 Experimental results and analysis

There are two subtasks in SIGHAN-8 topic-based Chinese microblog polarity classification

	All			Positive			Negative		
Team Name	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
TICS-dm	0.83	0.83	0.83	0.62	0.51	0.56	0.82	0.46	0.59
NEUDM2	0.74	0.74	0.74	0.31	0.08	0.13	0.44	0.08	0.13
LCYS_TEAM	0.72	0.64	0.68	0.26	0.05	0.09	0.40	0.10	0.16
HLT_HITSZ	0.68	0.68	0.68	0.21	0.40	0.28	0.45	0.60	0.52

Table 4: Performances in restricted resource subtask.

	All			Positive			Negative		
Team Name	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
TICS-dm	0.85	0.85	0.85	0.58	0.62	0.60	0.79	0.61	0.69
xk0	0.74	0.74	0.74	0.19	0.01	0.03	0.40	0.05	0.09
NEUDM1	0.74	0.74	0.74	0.26	0.11	0.16	0.46	0.33	0.38
HLT_HITSZ	0.71	0.71	0.71	0.24	0.41	0.30	0.51	0.54	0.53

Table 5: Performances in unrestricted resource subtask.

	All			Positive			Negative		
Approach	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Classifier 1	0.67	0.67	0.67	0.20	0.42	0.27	0.44	0.49	0.46
Classifier 2	0.60	0.60	0.60	0.18	0.61	0.28	0.42	0.67	0.52
Merging	0.71	0.71	0.71	0.24	0.41	0.30	0.51	0.54	0.53

Table 6: Performances by different classifiers in unrestricted resource subtask.

task: restricted resource and unrestricted resource subtasks.

Table 4 gives the performances in restricted resource subtask. The first column lists the name of participants who achieves higher macro average F1 values while out system is named as HLT_HITSZ. It is observed that our proposed approach achieves better performance on negative and positive categories, but obviously lower performance on neutral category. The good performance on the recall of minority classes showed the effectiveness of our consideration on imbalanced dataset training.

The achieved performances in the unrestricted resource subtask are listed in Table 5. Our system achieves about 3% of performance improvement on each category, respectively. It shows the contributions of extra training corpus and merging rules.

In order to validate the effectiveness of merging rules, the performances of Classifier 1 and Classifier 2 are evaluated, individually. The achieved performances are given in Table 6. It is observed that generally speaking, Classifier 1 achieves a higher classification precision because many features are coming from manually compiled sentiment-related lexicons. However, these features are limited to training data so that Classifier 1 achieved a lower recall. On the contrary, Classifier 2 may learn the representation features automatically from training data which is better for generalization. Thus, a good recall is achieved. Meanwhile, the achieved performances show that our joint model obtains better performances compared to two individual classifiers which indicate the effectiveness of our proposed joint classification strategy.

4 Conclusion

In this work, we propose a joint model for sentiment topic analysis on Chinese microblog messages. A word feature based SVM classifier and a SVM classifier using CNN-based paragraph representation features are developed, respectively. To overcome the limitation of each classifier, their classification outputs are merged to generate the final output while the merging rules are based on statistical analy-

sis on the performances on training dataset. Experimental results show that our proposed joint method achieves better sentiment classification performance over individual classifiers which show the effectiveness of the joint classifier strategy. In future, we intend to study the way to distinguish the subjective messages from objective messages for further improving the sentiment classification performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61370165,61203378), National 863 Program of China 2015AA015405, the Natural Science Foundation of Guangdong Province (No.S2013010014475), Shenzhen Development and Reform Commission Grant No.[2014]1507, Shenzhen Peacock Plan Research Grant KQCX20140521144507925 and Baidu Collaborate Research Funding.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Danushka Bollegala, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 132–141. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Minlie Huang, Borui Ye, Yichen Wang, Haiqiang Chen, Junjun Cheng, and Xiaoyan Zhu. 2014. New word detection for sentiment analysis. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 531–541, Baltimore, Maryland, June. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, October.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2267–2273.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 801–809.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011b. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–161. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. Association for Computational Linguistics.