NEUDM: A System for Topic-Based Message Polarity Classification

Yaqi Wang, Yang Wang, Shi Feng, Daling Wang, Yifei Zhang
Northeastern University, Shenyang, China
{wyqnumber1,wangyangdm}@gmail.com
{fengshi,wangdaling,zhangyifei}@ise.neu.edu.cn

Abstract

In this paper, we describe our system for the topic-based Chinese message polarity classification in SIGHAN 8 Task 2. Our system integrates two SVM classifiers which consist of LinearSVC and LibSVM to train the classification model and predict the results of Chinese message polarity in the restricted resource and the unrestricted resource, respectively. In order to assure our feature engineering effort on the task, we use some feature selection methods, such as LDA, word2vec, and sentiment lexicons including DLUT emotion ontology and NTUSD. Our system achieves the overall F1 score of 74.88% in the restricted evaluation and 74.43% in the unrestricted evaluation.

1 Introduction

With the development of social network, more and more people are actively sharing information with others and expressing their opinions and feelings on Chinese Weibo platform. Weibo has aggregated huge number of tweets that containing people's opinion about commercial products, celebrities, social event and so on. Therefore, mining people's sentiments expressed in tweets has attracted more and more attention for both research and industrial communities.

For the Chinese microblog, our task is to classify people's sentiments for a given topic as positive, negative, and neutral. Among the varieties of topics, people could express neutral, positive, and negative sentiments for them, respectively. If the topic information is ignored, it is difficult to obtain the correct sentiment for a specified target.

Topic-dependent features. The traditional learning-based methods for solving sentiment classification problem, such as (Go et al., 2009; Barbosa and Feng, 2010), basically followed (Pang et al., 2002), who utilized machine learning based classifiers for the sentiment classification of text. They worked in a topic-independent way: all the features have no relation with the topic. That is to say: the sentiment is decided no matter what the target is. Jiang et al. (2011) combined the target-independent features (content and lexicon) and target-dependent features (rules based on the dependency parsing results) together for tweet subjectivity and polarity classification.

Sparse vectors. The microblog usually has a length limitation, such as 140 characters. Therefore, the vectors formed by microblog data are extremely sparse, which sets obstacles for further classification algorithms.

To tackle these challenges, in this paper we first leverage the generative model LDA (Andrew Ng et al. 2003) to extract the top ranked topic words as topic-related features. Secondly, we count the number of positive and negative sentiment words through sentiment lexicon in the sentence and get the adjective word which only occur in the polarity sentences. Finally, we utilize the well-known deep learning word embedding tool word2vec[1] to find the top-k semantically similar words in the topic document to expand the feature representation. The used words in the word embedding tool word2vec both appear in a sentiment lexicon and the topic document. The component of feature vector can be described as follows.

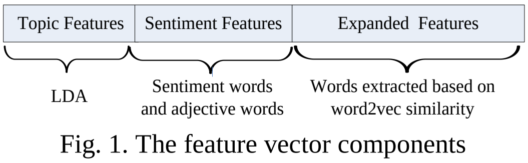
Topic Features
Sentiment Features
Expanded Features
LDA
Sentiment words and adjective words
Words extracted based on word2vec similarity

Fig. 1. The feature vector components

In Figure 1, the topic, sentiment and expanded features attempt to handle the topic-based sentiment analysis problem. The expanded features based on word2vec try to enrich the feature space and alleviate the sparse vector problem. Based on the feature vector, we can utilize off-the-shelf machine learning algorithms to train the topic-based sentiment classification model.

[1] http://code.google.com/p/word2vec/

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 185–189,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

2 Related Work

The traditional sentiment classification focuses on people's sentiment expressed in text. For example, whether a product review is positive or negative (Pang, Lee, & Vaithyanathan, 2002). Different from the traditional sentiment classification, topic-based classification is more challenging. The identification of topic-based sentiment needs to extract more information.

Jiang et al. (2011) proposed to improve target-dependent Twitter sentiment classification by 1) incorporating target-dependent features; and 2) taking related tweets into consideration. More specifically, they used two-step classification method to handle target-dependent twitter sentiment classification. They classified the tweets to the subjective and objective class, and then the subjective tweets are divided into positive and negative emotion class. Finally, they get the target-dependent twitter sentiment polarity.

Li Dong et al. (2014) proposed to the Adaptive Recursive Neural Network (AdaRNN) for target-dependent Twitter sentiment classification. AdaRNN adaptively propagated the sentiments of words to target depending on the context and syntactic relationships between them.

In this paper, we present machine learning based algorithms and deep learning system for SIGHAN 8 Task 2 which has restricted resource and unrestricted resource respectively.

3 System Overview

We use two-way classification framework which is used for the restricted resource and the unrestricted resource, respectively. Figure 2 illustrates the general framework of our system that includes the module of pre-processing, feature extraction, classifier training and predicting.

3.1 Data Pre-processing

Before the Chinese message is trained and predicted for the task, it needs to process the data so that the Chinese message can be split into words (tokenization). Meanwhile, it attaches more information to each word (part-of-speech tagging).

We adopt ICTCLAS2015[2] segmentation module, which is developed by Institute of Computing Technology, Chinese Academy of Science, to segment the given Chinese message including train and test data into words and proper part-of-speech (POS) tags. In the process, we delete the stop words, punctuation characters and other necessary processing. At last, we obtain the data for further feature extraction.

Chinese message	“魅族黄章叫板三星 Galaxy S6 也不过如此！ http://t.cn/RwHsCt6 @凤凰新闻客户端”
Bag of words	“魅族 黄章 叫板 三星 Galaxy S6 也 不过如此 ！ ”
Part-of-Speech features	“魅/w 族/ng 黄/nr1 章/n 叫/vi 板/ng 三星/nt Galaxy/n S6/n 也 /d 不过如此/vl ！ /wt”

Table 1: the example of ICTCLAS2015 segmentation result

3.2 Feature Extraction

The feature extraction plays very important role for the machine learning algorithms. A better feature extraction method can improve the prediction performance of the classifier, provide faster and more cost-effective classifier, and provide a better understanding of the underlying process that generated the data (Isabelle Guyon, Andre Elisseeff. 2003).

Due to the microblog can be split into many words and phrases, the overall Chinese microblog will be generate a rich set of features and may meet the curse of dimensionality problem. How to extract the appropriate features for topic-based sentiment classification is the key issue for the task. Our proposed approach of feature extraction includes:

Topic Features. Each microblog may be viewed as a mixture of various topic words. In order to obtain words and phrases which are relationship with the topic, we conduct LDA modeling for each topic collection and extract the top ranked words with higher topic generative probability.

Sentiment Features. We utilize the sentiment lexicons to select the sentiment features from Chinese microblog sentences. We calculated the number of positive emotion words and the number of negative emotion words in the Chinese microblog sentences, respectively. Then, we put the result into feature Set. The DLUT emotion ontology and NTUSD are chosen for the restricted resource setting of the SIGHAN task. For the unrestricted resource setting, we use our own sentiment lexicon. Besides the words in the lexicon, we also employ the POS tagging method to select the adjective words which only occur in the positive sentence and negative sentence as sentiment features.

Expanded features. To tackle the sparse problem of the short text in microblog, we em-

[2] www.nlpir.org

ploy the word embedding tool word2vec to enrich the feature representations. Given the topic documents *d*, *SF* is the feature set of *d*. The intersection set of sentiment lexicon and d is *ST* and $st \in ST$. Non-feature word $nw \notin SF$. Each word in the topic documents is represented as a vector based on word2vec (Dongwen Zhang et al., 2015) and then we calculate the cosine similarities between each sentiment and non-sentiment word pairs (*st*, *nw*). The *nw* with top-k similarity is added into final feature vector.

After feature selection, we utilize TFIDF method to calculate the weight of each feature in the vectors.

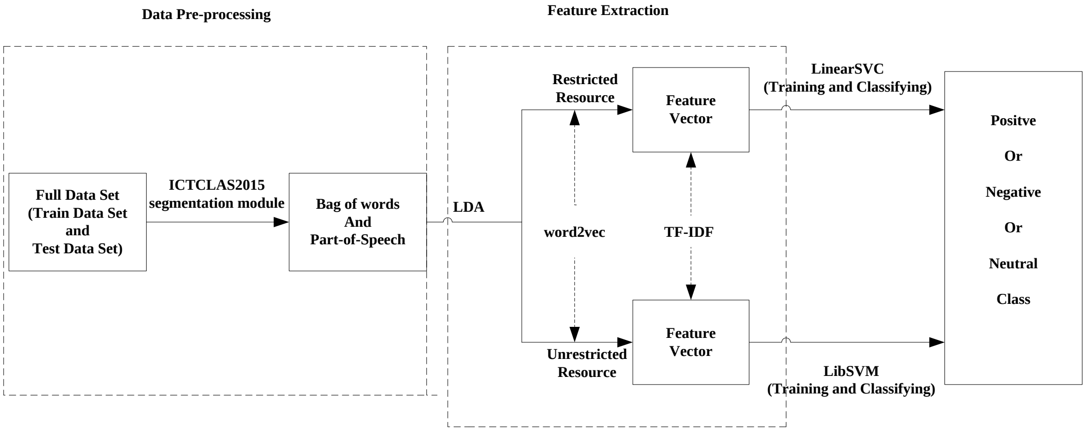

Fig. 2. The general framework of our system

3.3 Classifier

In this step, the extracted feature vectors are trained by a classifier to predict the sentiment polarity of the test data set. Lots of previous researches prove that Support Vector Machine shows substantial performance gains and is more robust in the work of sentiment classification compared with other state-of-art models (Pang et al., 2002; Tang, Tan, & Cheng, 2009). Due to this reason, SVM is adopted as the classification algorithm.

In the restricted resource, the SVM classifier is the linear support vector classification (LinearSVC) which is implemented in terms of Liblinear rather than LibSVM form the python based machine learning open source projects called Scikit-Learn[3]. In the unrestricted resource, the SVM classifier is the LibSVM which is implemented by C language. We search the best parameter c and g separately in the LibSVM for each topic dataset.

The implementation of Support Vector Classification is based on LibSVM. The fit time complexity is more than quadratic with the number of samples. Linear Support Vector Classification (LinearSVC) is similar to SVC with parameter kernel='linear'. It is implemented in terms of liblinear rather than LibSVM. The LinearSVC supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.

4 Experiments and Results

In this section, we explain details of the data and the general settings for the different experiments we conducted. We train and evaluate our classifier for restricted resource and unrestricted resource respectively, training and testing datasets provided by SIGHAN 8 Task 2.

4.1 Dataset

The train dataset is composed of five different topics and includes 4,905 Chinese microblogs. The test dataset is composed of twenty different topics and includes 19,469 Chinese microblogs. Each topic contains approximately 1,000 Chinese microblogs. But the ratio of subjective class to objective class is four to one and the ratio of positive class to negative class is one to one in the subjective class. The serious imbalanced data set has an adverse effect on classification results. So in order to keep the train data set balanced, we adopt the sampling strategy. We do not change the neutral class data and the sum of the positive class and the negative class is resampled to be an equal number of the neutral class.

4.2 Evaluation criteria

In SIGHAN 8 task 2, we evaluate the experimental results with Precision, Recall and F1 measure. These three classic values are utilized

[3] http://scikit-learn.org/stable

to measure the performance of positive, negative, neutral class respectively.

4.3 Classifier and Result

In the Section 3.2, the process of feature extraction has been done. We use the result of feature engineering into the classifier to train the classification model. In order to transform the each topic document in the train data set and test data set to vector matrix which complies with the input format of LinearSVC classifier in the restricted resource, we use the method of TF-IDF which is often used as a weighting factor in text mining and reflect how important a word is to a document in a collection or corpus.

If the adjectives appear only in the subjective sentences, the weight of the adjectives are set to 10. It means that the adjectives are more valuable. Meanwhile, we also make the input format complied with the LibSVM in the unrestricted resource and the parameter *c* and *g* of LibSVM in the unrestricted resource is set to *c*=8.0 and *g*=0.125.

At last, we use the classifiers which consist of LinearSVC in the restricted resource and LibSVM in the unrestricted resource to train the model and predict the label of the microblogs. Table 2 shows the result of the experiments.

The results in Table 2 show that the values of Precision, Recall, F1 measure is approximately equal to 0.74 in the restricted source and unrestricted source. We have achieve good performances in overall Precision, Recall and F1 measure. However, the values of Precision+, Recall+, F1+, Precision-, Recall-, F1- are not good. It means that the problem of imbalance data need to be better solved and we may further improve Topic-Based Chinese Message Polarity Classification task by adding more topic-related linguistic features.

Restricted								
Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
0.74883145	0.74883145	0.74883145	0.31879196	0.082465276	0.1310345	0.44460857	0.082715034	0.13948101
Unrestricted								
Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
0.74436283	0.74436283	0.74436283	0.17627119	0.045138888	0.071872845	0.40792078	0.05660896	0.09942085

Table 2: the results of our system

Restricted			
	Precision	Recall	F1
NEUDM2	**0.74883**	**0.74883**	**0.74883**
TICS-dm	0.83573884	0.83573884	0.83573884
LCYS_TEAM	0.7259232	0.7259232	0.7259232
Restricted			
	Precision	Recall	F1
NEUDM2	**0.74436283**	**0.74436283**	**0.74436283**
TICS-dm	0.85356206	0.85356206	0.85356206
xk0	0.74893427	0.74893427	0.74893427

Table 3: the compare of competition results

5 Discussion

After conducting a series of experiments, in this section, we discuss the effectiveness of our method. The compare of competition results in Table 3 show that the overall F1 score of our system is good. Firstly, the noise of Data is effective removed. Secondly, the quantity of feature is sufficient through the extracting of topic features, sentiment features and expanded features, respectively. Finally, LinearSVM is better than LibSVM and it trains faster and predicts more accurate in large-scale training set. As a result, the performance of the our system proposed method for Chinese microblog polarity classification is acceptable.

6 Conclusion and future work

Different from most of the conventional methods for subjective and objective classification, our research focuses on the topic-based polarity classification. In this paper, our system relied heavily on the topic features, sentiment and expanded sentiment features. These features assure the effect of our classifiers in this task. Our system, we can achieve medium score in the SIGHAN 8 task 2.

We have a lot of work ahead of us. In the future, we would like to find more topic-related linguistic features to add in the representation vectors. We would like to extract more structured information and composition unit existing in sentences for topic features in the future work.

7 Acknowledgements

This work is supported by the National Basic Research 973 Program of China under Grant No. 2011CB302200-G, the National Natural Science Foundation of China under Grant No.61370074, 61402091.

Reference

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet allocation". Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Alec Go, RichaBhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision.

Luciano Barbosa and Junlan Feng. 2010. Robust Setiment Detection on Twitter from Biased and Noisy Data. Coling 2010.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT ’11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Isabelle Guyon, Andre Elisseeff. 2003. An Introdution to Variable and Feature Selection.

Dongwen Zhang, Hua Xu, Zengcai Su, Yunfeng Xu. 2015. Chinese comments sentiment classification based on word2vec andSVM^{pref}.

Pang, B., Lee, L., &Vaithyanathan, S. (2002). Senment classification using machine learning techniques.In Proceedings of the conference on empirical methods in natural language processing (pp. 79–86). Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification.