Chinese Micro-blog Sentiment Analysis Based on SVM and Complex Phrasing

Yang Fuping[1], Huang Zhiyong[2]

Chongqing University of Posts and Telecommunications, Computer science and technology school， Chongqing 404100,China

[1]yangfp@cqupt.edu.cn, [2]287590296@qq.com

Keywords: micro-blog, Sentiment analysis, SVM, Naïve Bayes, Complex phrasing

Abstract. Text sentiment analysis technology is a hot topic recently. As a short text, there is a feature of using complex sentences to express the author's true views and complex emotional tendencies in micro-blogs. In current researches on sentiment classification based on machine learning, few of them focus on complex sentences. This paper proposed a sentiment analysis method based on SVM and complex phrasing classifier, and made a full analysis of structural features of Chinese conditional sentences, transition sentences and multiple negative sentences, which were taken as text features. A variety of different combinations of features were chosen, including emotional words, speech, negative words, the degree of adverbs and punctuation, etc., to optimize the results of sentiment analysis through multiple sets of experiments. The experiments show that when we choose the combinations of features of emotional words, part of speech and complex sentence patterns, this method improved the accuracy of sentiment classification compared to the common method.

Introduction

Recently, sentiment analysis has become a hot topic in the field of natural language processing. The previous works have covered a wide range of tasks, including polar classification, extraction of view[1] and opinions origins. Depending on the size of text, sentiment analysis can be divided into several levels including chapters level[2], sentences level[3] and words level[4]. According to the latest data from Sina, registered users in micro-blog has outnumbered 500 million, and 200 million active users monthly. Micro-blog is continuously influencing the life of numerous netizens with updating micro-blog has became one of the habits to more and more people.

At present, the domestic sentiment analysis of Chinese micro-blog are based on machine learning with the feature normally choosing to use the knowledge from statistics, which cannot express the meaning of the context comprehensively and lack of the effective analysis on complex phrasing. This paper studies the structural features of Condition sentences and Turning sentences, which proposes new rule for sentiment analysis on complex phrasing mentioned above to take it as text feature combining with a variety of features. The experimental results have proved that the proposed method brings obvious effect on improving the accuracy of sentiment classification.

Relative Work

The research idea can be concluded into two types, one is based on sentimental dictionary, the other is based on machine learning. This paper will set forth from the two sides.

The method based on emotional dictionary

This idea needs to collect words with emotional colors manually, then classify them into positives and negatives, finally build emotional dictionary. When judging emotional tendency, the

This work is supported by Scientific and Technological Research Program of Chongqing Municipal Education Commission（Grant No. KJ130532）

first step is to calculate the numbers and its sentimental value, then calculate the influence on sentimental value by adverb and negatives, lastly, judge the sentimental tendency according to the different value. People including Xiaodong Chan[5] take the emotional words in micro-blog as feature to calculate emotional tendency from emotional value, the accuracy is 74.2%. But this rule-and-emotional dictionary-based method has great limitations. Firstly, it is hard to collect complete emotional dictionary; secondly, a lot buzzwords emerged in Internet; besides, there are ambiguity in some words in different context or filed, the emotional polarity cannot for sure.

The method based on machine learning

Davidiy et al.[6], who references the thought of K- nearest neighbor, classifies tweet into several kind of emotional types through emoticons and hashtag; Hassan et al.[7] judge Newsgroups' information polarity based on HMM as well as combined dependency parsing and syntactic character; Turney et al.[8] proposed a method based on Mutual Information to analyze emotional polarity of specific phrases, then judge commented emotional tendencies. Dongwen Zhang et al.[9] use Word2vec to choose semantic features through semantic relationships between words, and achieved good effects on sentiment classification. Meena et al.[10] focuses on the emotional impact of conjunctions for sentence, but this method cannot adapt to different areas. Socher et al.[11] obtain some sentences' vector space based on semi-supervised recursive automatic coding machine, and then use recursive analysis to express the whole sentences, predict emotional tendencies.

Sentimental Sources Building

Pre-treatment of data

The url data, @users, #topic# and websites in micro-blog excluding the viewpoints of users, and likely become distraction of words classification and part-of-speech tagging with side effects. Screen information above before words classification, and use jieba words classification tool to classify and tag character, then use stop words. Use the HIT stop words form to process, discard the conditional words and turning words etc.

Sentimental dictionary

The emotional dictionary in this paper combine HowNet[12] and NTSUSD, eliminate repeated parts, unfamiliar words and vague emotional words; at the same time supply the evaluation buzzwords from Sogou.com, at the last, get 9063 positive words, 10,270 words a negative words, 19333 emotional words in total.

Dictionary of negatives

Negatives is one kind of adverb to change polarity of emotional words. This paper has collected some common negatives，a total of 32.

The form of conditional words

Conditional words belong to conjunction, the conditional words would weaken the strength of emotional words. E.g.: This cell phone will be perfect if it has better tone quality." when there is "if", the strength of perfect is weaken. There are some relative conditional words, such as "if", "once" et al. a total of 25.

The form of turning words

Turing words are one kind of conjunction, Di Peng et al.[13] classifies turning conjunction into two types: one is concession words like "although"、"though" which stands for positive turning; and other kind like "but" to show the turning relation directly with the real feeling after the conjunction. This text focuses on the second type, which in total 14words.

The form of degree adverbs

The degree adverbs strengthen or weaken the strength of emotional words. Some degree words are collected by different intensity value between 0.5 to 2.0. For example: the intensity of "the slightest" is 0.25, and 1.25 for "more", 2 for "absolute", etc.

A Variety of complex sentences

Introduction of complex sentences

According to the classification standard of modern Chinese, one sentence usually summarized as a Simple Sentence or Complex Sentence. Simple Sentence contains only one “subject + verb” structure which comes up with words or phrases to show the simple meaning; while complex sentence has pause and associated word compared to Simple Sentence, generally, it contains two or more independent “subject + verb” structures with complex meaning.

Conditional Sentences

Conditional sentences generally raise assumption by using words like “if” in conditional clauses, then express the result in the main clause. For example: "This movie is so so with complex plot. If cut the ending, this movie would be perfect.” If calculate emotional intensity value by traditional feature extraction, micro-blog is the positive feeling. However, if we think from natural language, then its negative. Taking all into consideration, this paper research conditional sentences on emotional tendency analysis.

Transition Sentences

According to have transition points or not in transition sentences, there are two types: narrow transition including conjunctions and general transition with no obvious conjunctions. This paper focuses on the narrow transition, as the second category conjunction of the transition in 2.5.

Experiment and results analysis

Experimental data

This paper collects corpus supported from COAE2013 task4、 NLPCC2012 and NLPCC2014, and get positive micro-blogs and negative micro-blogs 5,000 pieces individually. Then, take each 4,000 pieces as training set and the remaining 2,000 as a test set.

Sentiment analysis based on SVM

This paper studies on commendatory and derogatory classification of subjective information. Firstly, pre-process micro-blog and extract text feature, then test various combinations of features in SVM classification models.

Features choosing

This paper selected seven text features, including emotional words, the conjunction and conditions words and so on. Table 1 lists all types of features and meanings thereof.

Table 1: types of features and meanings thereof

Type of features	Meaning
Speech	The number of verb, nouns, adjectives and adverbs
Emotional Words	The number, intensity value and emotional value of positive or negative words
Negative Words	Mathematics take over 2 of the appearing time of negative words
Degree adverbs	Is there any degree adverb before emotional words
Conjunction	Is there any conjunction word before emotional words
Conditional Word	Is there any conditional word before emotional words
Punctuation	The number of exclamation and question marks

Features extraction

After pre-setting, classifying and part-of-speech tagging the training corpus, extract seven features in table 5 to train in SVM model, then, test with test set corpus. Count the times positive words appearing before emotional words, take off 2, if the result is 1, then the polarity take the opposite one; if the result is 0, the polarity stays still. Emotional intensity changes as degree adverb showing before emotional words. When conditional words or conjunctions appears before emotional words, it will reduce or rise emotional intensity. Through random corpus analysis, the strength of the emotional words of enhanced 1.5 times if conjunction appear before assumptions emotional words,; when

conditional words presence at the beginning of a sentence, the intensity of word reduce to 25%. This article takes emotion words as features, which refers to the sums of emotional intensity value in the text calculated from the number of the presence that negative words, conjunctions, and conditional words before the emotion word. The emotional value calculation method as shown in formula 1:

$$sentiValue = \sum_{i=1}^{n} (x_1 \mid x_2) * [f_1(neg\%2) * f_2(adv) * f_3(w_i)]. \quad (1)$$

In formula 1, x_1 represents the improving coefficient from conjunction to the intensity emotional words, which is 1.5; x_2 represents the reducing coefficient from conditional words to emotional words, which is 0.25, besides, there will be only one conjunction or conditional words in one clause. $f_1(neg\%2)$ represents mathematics take over 2 of negative words (neg), then function value is -1 or 1 when the result is 1 or 0. $f_2(adv)$ presents the intensity of degree adverbs (adv). Lastly, $f_3(w_i)$ stands for the polarity of emotional words w_i with the form of -1 or 1.

Feature combination and experimental design

The features selected from the experiments in this paper including emotional words, the degree adverb, the conjunction, conditional word and punctuations. In order to find the best combination of features, multiple sets of experiments has been taken in the SVM model: at the beginning, taking emotional words as the only feature to test, and then add characteristics including addition of speech, the conjunction, the conditional word and punctuation.

There are two Baseline methods for comparative experiment, respectively based on l SVM model and Naïve Bayes model using the seventh features combination in experiment 1. The emotional value from the emotional words' features used in Baseline is just the sum of emotional intensity in text instead of considering the effects from conjunctions and conditional words. Experiment on the same experimental data.

Experimental results

The experiments use precision recall and macro F1 to judge the result. The definitions are as follows:

$$precision = \frac{System.Correct}{System.Output}. \quad (2)$$

$$recall = \frac{System.Correct}{HumanLabeled}. \quad (3)$$

$$F1 = \frac{2 * precision * recall}{precision + recall}. \quad (4)$$

Wherein, System.Correct represents the matching number of micro-blogs between emotional tendency results and artificial labels; System.Output represents the total number of micro-blogs concluded by emotional tendency; Human.labeled represent the number of micro-blogs with artificial labels. Table 2 lists the results of various combinations of features.

Table 2. The experimental results of different features combination

Experiment1	Feature combination	Accuracy/%
1	Emotional words	84.96
2	Emotional words + Speech	87.14
3	Emotional words + Speech + Negative words	88.38
4	Emotional words + Speech + Negative words + degree adverbs	87.19
5	Emotional words + Speech + Negative words + punctuation	87.22
6	Emotional words + Speech + Negative words + Conjunction	90.23
7	Emotional words + Speech + Negative words + Conjunction + Conditional word	91.23

The experimental result comparison of method from this paper and Baseline is shown in Fig. 1.

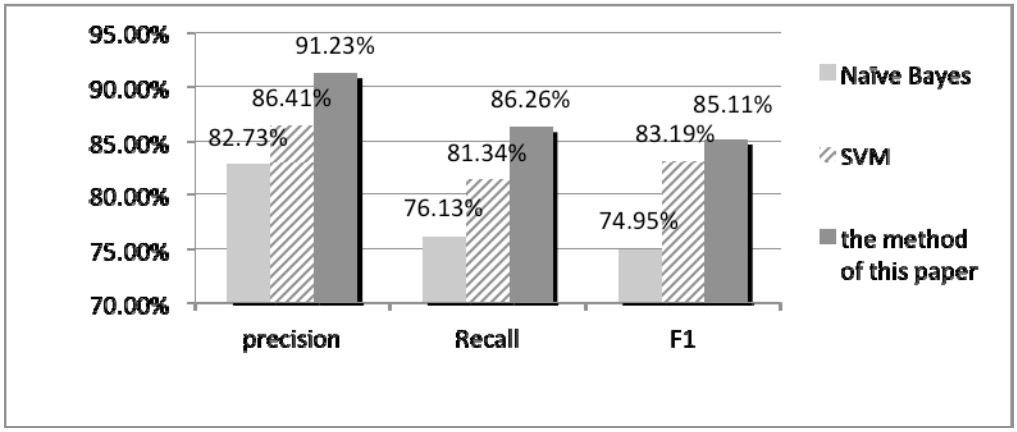

Fig. 1 The experimental result comparison of method from this paper and Baseline

Results analysis

Table 2 shows that using emotional words, part of speech, the conjunction and conditional word as the word combinations of features in SVM model reaches the best accuracy rate, 91.23%, among which emotional words has the biggest effect 87.52% separately, conjunctions 2.80%, condition words 0.89%. Since the effect of emotional words have taken conjunctions and conditional words as one part, the effect of the above two features is not so apparent. After the integration of degree adverbs and punctuation features, the accuracy of this experiment slightly reduce instead, indicating that the use of these two features in SVM model is not helpful.

Fig. 1 has shown that SVM model with adding transition sentences or conditional sentences features has higher accuracy and recall rate compared to common ones. The accuracy of SVM model accuracy and Naïve Bayes model for Complex Sentences emotional classification is not very high. Because of the transition sentences often have multiple co-existing emotion, and emotional words after conjunctions express the author's true feelings; conditional sentences is real or fake to past or the future assumption from general logic deduction, which weakens the strength of the emotion words in a large extent. The traditional sentimental classification model ignores the characteristics of Chinese sentence structure, making it difficult to identify the emotional in Complex Sentences. In daily life, people often use Complex Sentences to express themselves, besides, transition sentences and conditional sentences have a greater impact on the traditional text emotion classification model, thereafter, it is necessary to deal with Complex Sentences.

Summary

This paper presents a method based on SVM and sentiment analysis of Complex Sentences, and achieves good results on corpus from COAE2013, NLPCC2012 and NLPCC2014. Experiment 1 finds out that the best result of SVM model is when using emotional word, part of speech, negative words, the conjunction and conditional words, while the degree of adverbs and punctuation feature reduces the effect of sentiment classification; experiment 2 shows that the fusion complex features in SVM model sentence can effectively improve the accuracy of sentiment classification, compared with the method based on SVM model before, the classification accuracy improved 4.82%, reaching 91.32%.

References

[1] Pang B, Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts[J]. Proceedings of Acl, 2004:271--278.

[2] BALAHUR A.STEINBERGER R. KABADJOV M ., et al. Sentiment anaylysis in the news[J]. Infrared Physics and Technology,2010.65:94-102.

[3] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter Sentiment Classification[J]. Proceedings of Annual Meeting of the Association for Computational Linguistics Human Language Technologies, 2011, 1:151-160.

[4] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2):267-307.

[5] Cheng Xiaodong. Research on Sentiment Dictionary based Emotional Tendency Analysis of Chinese MicroBlog [D]. Wuhan: Huazhong University of Science and Technology,2012

[6] Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys.[C]// International Conference on Computational Linguistics: Posters. 2010:241-249.

[7] Hassan A, Qazvinian V, Radev D. What's with the Attitude? Identifying Sentences with Attitude in Online Discussions[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010:1245-1255.

[8] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVMperf[J]. Expert Systems with Applications, 2015, 42(4):1857–1863.

[9] Turney P D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[J]. Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002:417--424.

[10] Meena A, Prabhakar T V. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2007:573-580.

[11] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011:151-161.

[12] Dong Zhendong, Dong Qiang. Hownet[DB/OL].[2012-09-15]. http://www/keenage.com/. (in Chinese).

[13] DI Peng; LI Ai-ping; DUAN Li-guo. Text sentiment polarity analysis based on tarnsition sentence[J] // Computer Engineering and Design, 2014, 12(12):4289-4295