Emotional Intensity Calculation for Topical Microblog Based on Comments' Emotion

Jia Zheng, Zhengping Jin

School of State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;

zhengjia_bupt@163.com, zhpjin@bupt.edu.cn

Keywords: Emotional Intensity Analysis, Topical Microblog, Authentication Type.

Abstract. Emotional intensity analysis is widely applied in Twitter and Microblog. Moreover, the latest approach to calculate emotional intensity of Microblog is taking Microblog as an entirety and ignoring the comments. Because most of the topical Microblog has the topic comments which include emotional intensity, it is important to take topic comments into consideration. In this paper, we solve the problem by two steps. Firstly, we construct the sentiment lexicon, and then process comments of the topic to improve accurate of emotional intensity. Secondly, we use Authentication type of Microblog to optimize emotional intensity result. The experimental results demonstrate our method to calculate Chinese Microblog emotional intensity more effectively.

Introduction

Similar to the Twitter, Microblog is also the most popular in China. Microblog encourages people to throw their opinions and share their mood. Microblog and Twitter is different. First of all, Microblog users can publish their messages up to 140 Chinese characters which may contain several sentences, but 140 English characters usually just contain one sentence. However, Microblog 140 characters limit, and most users care about the weather, life, movies, feelings, emotions and other everyday topics, and there are a large number did not isolate users to communicate with people, microblog not have a complete timeline, space axis, the subject and style. Despite the different, Microblog had attracted more and more people study.

Microblog is becoming more and more important in people's life, an increasing number of people express a view of themselves, others or society in the Microblog. Take Sina Microblog, the number of registered accounts had climbed above 500 million up to now, more than 100 million Microblog content generated every day. In fact, as the representative of social networks Microblog has become one of the most important media in China. The Microblog itself has a description, forwarding, discussion and other functions which can cause a hot topic of high rate of comments and high rate of forwarding and lead to a rapid expansion of Internet public opinion if a major event caused the attention of netizen. If the government can't find the positive or negative impact on the people of the incident in time, public opinion is likely to embark on extreme and difficult to control.

In this paper, we focus on topical Sina Microblog sentiment classification. The topic in Microblog is starting with symbol "#" and ending with symbol "#", most of the topics are hot issue, the topic has provoked a wave of discussions on Microblog. For emergencies always can become a hot topic, some people are positive for the hot issue, others think this hot issue is so bad, and the rest think that it is not a big deal. Our work is classifying Microblog of topic into positive, negative or neutral and calculating the degree of negative impact and positive impact. So the Public opinion monitoring related departments or government can obtain people's sentiment to the hot issue and control the negative impact of the hot issue and take effective measures to correct public opinion guidance.

The remainder of this paper is structured as follows. In Section 2, we briefly summarize related work. Section 3 gives an overview of our method, and we construct sentiment lexicon and compute the emotional intensity of the topic and optimize emotional intensity of the topic. Experimental results are reported in Section 4 and Section 5 concludes our work.

Related Work

Sentiment classification is a hot topic in Natural Language Processing. Most existing work focuses on unsupervised method or supervised method for classification. After Microblog appear, increasing scholars study sentiment classification of Microblog, and their methods can mainly divided into lexicon-based method and machine learning method.

For the lexicon-based method, the challenge is how to construct a comprehensive sentiment lexicon and calculate the sentiment polarity of Microblog.

1. Tong use manual method to establish a emotional dictionary, this dictionary mainly aims at the movie critics [1]. The disadvantage of this method is that the applicable surface of the emotional dictionary is relatively simple, general can be aimed at a particular field, so by using this method to analyze the target all need each building the corresponding emotional dictionary.

2. CUI et al., construct classify based on the out-of-vocabulary lexicon for Microblog messages [2].

3. Such as HOU et al., construct a coarse grain sentiment lexicon and design rules to classify Microblog messages [3].

4. Lin et al., use LSPM (Latent Sentence Perspective Model) model to determine the point of view of the statement and its five categories of emotional intensity to judge [4].But because this method is ignored in the process of the results of a gradual process of emotional continuity, leading to the learning model is not accurate enough and have a great impact on the classification accuracy [5].

For the machine learning method

5. Hu et al., train models totally from unlabeled data and sentiment lexicons. The disadvantage of this method, without supervision from labeled data, cannot provide good performance [6].

6. Such as Jiang et al., propose the target-dependent feature and context-aware feature for tweets classification; and propose a series rules to extraction dependency parser features for Twitter messages classification [7]. The method is useful when the Twitter messages have the target, but the messages of Twitter messages are oral expression and grammar express style is free, so the rule extraction of dependency parser features is very hard design.

As mentioned in above, the latest approach to calculate emotional intensity of Microblog is taking Microblog as an entirety and ignoring is comments. In this paper, we compute the emotional intensity of Microblog by using information-publish microblog and commentary microblog, and optimize emotional intensity result by using Authentication type of Microblog.

Approach Overview

A. Construction of Sentiment Lexicon

The collection of Sentiment Lexicon is to choose the right corpus to do the pretreatment for labeling of corpus in advance. The method of corpus selection is related to the coverage of corpus, the so-called coverage is the distribution or distribution of the corpus in different fields. Different fields usually refers to the style axis (reflect stylistic characteristics), subject axis (reflect knowledge characteristics), space axis (reflect regional characteristic) and time axis (reflect the features of The Times) consists of four dimensional model.

This article constructs its own Sentiment Lexicon. When it encounter unfamiliar words, it can add to the corpus through the self-learning algorithm. Multi-classification is the text of each emotional intensity level as a category, thus creating the classifier on its classification. Generally speaking, the emotional intensity of the word can be divided into strongly comstockery, generally comstockery, objective, and general praise, and strongly praise the five categories. Lin et al., use LSPM (Latent Sentence Perspective Model) model to determine the point of view of the statement and its five categories of emotional intensity to judge [8].

B. Computing the Emotional Intensity of the Topic

After build the Sentiment Lexicon, we should compute the Emotional Intensity of hot issues. Before computing, we study the sentiment express style of Chinese Microblog. The Chinese Microblog usually have several sentences (sentences means split by full stop, not the comma symbol), and sentences contains difference target with difference sentiment polarity. So we classify Microblog

by the sentence level instead of Microblog level. Basically the sentences of Microblog can classification to two categories, subjectivity sentences and objectivity sentences. Obvious, the objectivity sentences is neutral sentiment, the subjectivity sentences contains sentiment which can classify to positive sentiment and negative sentiment. The subjective sentences are containing user sentiment or speculate, the objective sentences are factual description which not contains user sentiment.

Because the comments are very complex, and a lot of comments is not coherent, invalid comment. If the calculation of these invalid comment, not only a waste of computing resources and reduce the accuracy of the emotional intensity, so it is necessary to review the following pretreatment operation：For comments set C

(1) Remove objectivity sentences.

(2) Divide the comments like “comment//@user 1: comment 1//@user 2: comment 2//...../@User e: comment e” into several comments like “comment”, “comment 1”, “comment 2”, ... ,“comment e”.

(3) Remove repetitive comments.

(4) Remove comments less than 3 words.

(5) Remove comments only containing numbers, letters, special symbols and punctuations.

The next step is computing the emotional intensity of the microblog text. Emotional value is text emotional orientation for positive mood, emotion negative value for representing text emotional orientation for negative emotion, emotional value of 0 indicates no emotion or emotion neutrality. Emotion value is higher, the greater the degree that positive emotion; emotion value is lower, that the greater the degree of negative mood. The basic idea of computing method is as follows:

- A microblog text be read into the of ICTCLAS word software. The text is divided into a number of words a1, a2,..., an, in accordance with the computing annotation method.
- a1 as the attitude of words in the corpus of standard word search, return to the attitude of words in the corpus = a1 weights v1; less than if the search returns a null value, ie v1 = null.
- a2, a3 ..., an as a standard word, repeat steps (2). Get v2, v3, vn.
- V1, V2,..., Vn can be obtained by adding the V, mean k1=v/n as a microblog text a the emotion value.
- Read the next Microblog b repeat the steps to get emotional value k2. Similarly cycle that k3, k4,….Kn.

$$Text(mood) = \frac{\sum K_n}{n}. \tag{1}$$

But information-publish microblogs and commentary microblogs are difference. Because the post is more direct to guide public opinion, so it's not directly superimposed on their emotional value to reflect the emotional intensity of the topic. In this paper：

$$Topic(mood) = \frac{\sum K_c}{n_c} * 0.01 + \frac{\sum K_p}{n_p} \tag{2}$$

C. Emotional Intensity of the Topic Optimized

The above calculation method ignores the identity of the post. For example, if the identity of the publisher is the government agency or media certification, its blog is more powerful and more persuasive than personal authentication. The Microblog user level is shown in the table below.

Table 1 Authentication type

Authentication type	privilege	authentication	difficulty
Government certification	Identity recognition	Blue Icon	Easy
Enterprise certification	Search priority	Blue Icon	medium
Agency group	Privilege of speaking	Blue Icon	diffcult

Under the guidance of public opinion and emotional intensity of topic Microblog, the authentication level of different Microblog user is not the same. In order to calculate the emotional intensity of topic Microblog more scientific, we use a weighted method to give different weights to different kinds of users in this paper. For example the weight of the government certification users is

higher than the individual users. Then calculate the emotional intensity of each topic Microblog values.

Experiments

In order to verify the effectiveness of microblog's emotional intensity incorporation comments, experiments are conducted from two aspects:

(1) Evaluation of effectiveness of introduction comments.

(2) Emotional Intensity of the Topic Optimized.

Because now there is no universal microblog dataset in China, we collect Sina Microblog data to do related experimental analysis. The dataset contains 20 topics, with 5000+ information-publish microblogs and 31,675 commentary microblogs. Among them, 2,023 microblogs and 3,416 sentences in 20 topics have been annotat- ed with subjectivity and polarity.

A. Evaluation of effectiveness of introduction comments

First we have to deal with the comments. In the process of the pre-processing of the comments, we find that the pre-processing can improve the accuracy of the emotional intensity of the topic. We choose three typical microblogs respectively from 20 topics. The processing results of microblog comments are shown in the table below.

Table 2 Pre-processing of the Comments

Comments of topic	No processing	Processing objective	Processing Irrelevant
1	1812(461)	954(411)	396(359)
2	1955(695)	1140(632)	687(615)

From TABLE 2, we can see that the amount of comments is reducing with the processing of comment step by step. At the same time the effective comments ratios are rising sharply before and after using the processing of comment. Take the first microblog in TABLE I for example, before and after using the processing of comment, its amount of comments declines from 1812 to 396, but the effective comments ratio rises from 25.4% (461/1812) to 90.7% (359/396). It proves the effectiveness of the processing of comment.

Therefore, whether introducing comment is effective is estimated by the following standards: (1) whether the comment contains the keywords in the microblog text. (2) if the comment doesn't contain the keywords in the microblog text, the estimate is conducted by whether the comment's content is associated with microblog content's extension, influence and further development. The emotional intensity value of the topic is shown in the table below.

Table 3 Emotional Intensity Comparing

topic	no comments	comments（no processing）	comments（processing）
1	0.453	0.672	0.713
2	-0.562	-0.597	-0.695
3	0.378	0.479	0.573

B、Emotional Intensity of the Topic Optimized.

The dataset of the topic, 56.20% Microblog texts are Authentication type. That means our Authentication type optimization is potentially useful for most of the topics. Experiment select 2 representative topic which have at least 1500 Authentication types. We use the train model to compute the value of emotional intensity of topic, and then we use Authentication type to optimize the computing result. The optimization result is shown in the table below.

Table 4 Optimization Result

topic	Initial state	Introduce grade classification
1	0.713	0.768
2	-0.695	-0.736

From TABLE 4, we can see that the value of emotional intensity of the topic microblog is better than before. Accuracy, evaluation of emotional intensity of the topic is significantly improved. It proves the effectiveness of Authentication type optimization.

Conclusions

In this paper, we propose a method incorporation comments to compute the emotional intensity of topical Microblog. According to the experimental results, our methods outperform previous method by comparing accuracy. The experiments result show the processing of comments is necessary, Authentication type of Microblog optimization is helpful to improve the accuracy.

We set five categories of emotional intensity to judge which is strongly comstockery, generally comstockery, objective, and general praise, and strongly praise. But this method is ignored in the process of the results of a gradual process of emotional continuity, leading to the learning model is not accurate enough and have a great impact on the classification accuracy.

This paper considers Authentication type of topic Microblog and uses them to calculate the emotional intensity of the topic Microblog. But we ignored the comments in the number of praises, so the number of praises should be included in the calculation in the next step of research. And there are some comments on the topic Mircoblog that has no relationship with the comments which is invalid comments. These comments also affect the results of the calculation of emotional intensity, so how to effectively remove these invalid comments is the next research content.

Acknowledgements

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

References

[1] Tong R M. An operational system for detecting and tracking opinions in on-line discussion [C]//SIGIR Workshop on Operational Text Classification. NY, USA, 2001: 1-6.

[2] Cui, A., Zhang, H., Liu, Zhang, M., Ma S. "Lexicon-based Sentiment Analysis on Topical Chinese Microblog Messages. " 2012. NLP&CC 2012.

[3] Hou Min, Teng YongLin, Li Xueyan, Chen Yuqi, Zheng Shuangmei, Hou mingwu, and Zhou Hongzhao. "Study on the Linguistic Features of the Topic-oriented Microblog and Stategy of its Sentiment Analysis. " 2012. NLP&CC 2012.

[4] Lin W H, Wilson T, Wiebe J, et al Which side are you on Identifying perspectives at the document and sentence levels [C]//Proc. of the 10th Conf. on Computational Natural Language Learning (CoNLL-X), NY, USA, 2006: 109-116.

[5] WangGen, zhao jun. Based on multiple redundancy markers of sentence analysis CRFs emotional [J]. Journal of Chinese information 2007, 21 (5) : 51-55

[6] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In WWW, 2013.

[7] Long Jiang, Mo Yu, Ming zhou and Xiaohua Liu. "Target dependet Twitter Sentiment Classification. " In Proc. ACL-HLT, pages 151-160, 2011.

[8] Lin W H, Wilson T, Wiebe J, et al Which side are you on? Identifying perspectives at the document and sentence levels [C]//Proc. of the 10th Conf. on Computational Natural Language Learning (CoNLL-X), NY, USA, 2006: 109-116.