Research on the Impacts of Quantitative Factors on Sentimental Classification of Weibo of Different Topics

Ruoxi Zhang[1, a]

[1]International School , Beijing University of Posts and Telecommunications, Beijing 102209, China.

[a]zhangrx@bupt.edu.cn

Keywords: Sentimental classification, training set, quantitative factors, Weibo.

Abstract. Different numbers of terms and texts of training set are involved to optimise the performance of sentiment classification for Weibo repost. The experiment utilises CHI-square test to extract terms and uses support vector machine (SVM) to classify the different sentimental categories. By measuring the performance by F1-score for different training sets, the result illustrates the impacts of quantitative factors on the performance and the differences of the impacts between particular topics.

Introduction

The sentiment analysis on Weibo text illustrates the sentimental attitudes of users toward it in network communication, which is valuable in that it could assist public opinion monitoring and commercial promotion. This prevailingly involves machine learning techniques to implement sentiment classification. Liu Z and Liu L discussed the performance of different combinations of varied machine learning methods, such as support vector machine and Naive Bayes, term extracting methods and tern weighting methods [1]. Zhang Q, Zhang L, Dong S and Tan J pointed out the importance of training set, which have direct influence on the performance of the classification [2]. Wu D, Ye Q, Ye N, Zhang X, and Wu B focused on optimising the performance by selecting proper training set [3]. Li X, Cao H and Huang L discussed three quantitative factors that affect the training set: the number of terms, the number of texts and the number of categories [4]. However, these researches didn't make subdivided discussion for different topics, and there may be differences to optimise the classification performance for diverse topics.

Considered that the number of categories is fixed in sentiment analysis, only the number of terms and that of texts will be considered for two non-identical topic to discuss their impacts to the performance of sentimental classification.

Methodology

Data collection. The Weibo repost data is collected from Sina Weibo by reptile tool *Bazhuayu*, which covers 7334 pieces in total, comprising for environmental topic: 1830 sentimentally positive reposts and 2011 negative ones; for business promotion topic: 1725 positive, 1768 negative. Qualitative sentiments are marked manually.

Data preprocessing. Firstly, the non-sentimental contents, such as topic name between two hashes (#) and the interaction of users (@) are rejected. Then the sentences in each repost are split by *ICTCLAS*. Customised dictionary is added to adapt the popular network neologisms..

Sentimental terms extracting. CHI-square test is used to extract the sentimental terms to be used to generate the eigenvector for each repost. CHI-square test measures the dependency between a term and a particular sentiment. A higher CHI value indicates a closer relationship.

$$CHI(t,c_i)=\frac{N\times(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \tag{1}$$

The data collected is used as the corpus to calculate the CHI value of each term. Threshold values are introduced to remove the terms whose CHI value is too high or too low.

Table 1 CHI-square parameter explanation

Parameter	Explanation
t	Sentimental term.
c_i	Category *i*.
N	The total number of the reposts.
N_{11}	The number of the reposts that contains term *t* and belongs to category *i*.
N_{10}	The number of the reposts that contains term *t* and doesn't belongs to category *i*.
N_{01}	The number of the reposts that doesn't contain term *t* and belongs to category *i*.
N_{00}	The number of the reposts that doesn't contain term *t* and doesn't belong to category *i*.

Eigenvector calculation. The eigenvector is calculated for each repost to represent it and could be later used in classification. Each repost is compared with the all the extracted terms. The frequency of each term in a particular repost is recorded, represented as a multi-dimensional vector. This vector is then normalised to be an eigenvector to signify the repost.

Sentiment classification. Support vector machine (SVM) is used to implement sentiment classification. SVM, based on the structural risk minimisation (SRM), searches an optimal hyperplane to divide one tuple from others, which realises the classification of different sentiments.

The sets of eigenvectors of training data of different sizes for each topic is used to train SVM models as the classifiers. In the experiments, an open source package *libsvm* is used to implement the classification work. After training the model, the testing data is inputed to the model to be classified. Each output F1-measure is used to evaluate the performance for one experiment. A higher F-score signifies the better performance.

$$F1\text{-}measure=\frac{2\times Recall\times Precision}{(Recall+Precision)} \tag{2}$$

Performance comparison. In each experiment, both training set and testing are sampled randomly. The size of testing set for each topic is constantly 500, while the one of training set is respectively 500, 1000, 1500, 2000, 2500, measuring the influence of text number. In addition, different quantities of terms are used to measure the impact of term number. Each F1-measure is recorded to represent the performance.

Results

Taking the environmental topic as an example, when the number of text in the training set is fixed (1000), the value of F1-measure increases with the growing number of terms, and it reaches the peak (0.693) when the term number is 2600. After that, it begins to decrease smoothly. For the topic of business promotion, the tendency of F1-measure is similar to that of the environment topic. When the text number is 1500 in business topic, F1-measure keeps climbing until it gets to the top (0.0663) at term number of 2200 and then starts to decline with fluctuations.

In general, the F1-measures of both topics increase with the growth of the text numbers and reach the maximum value at a particular term number.

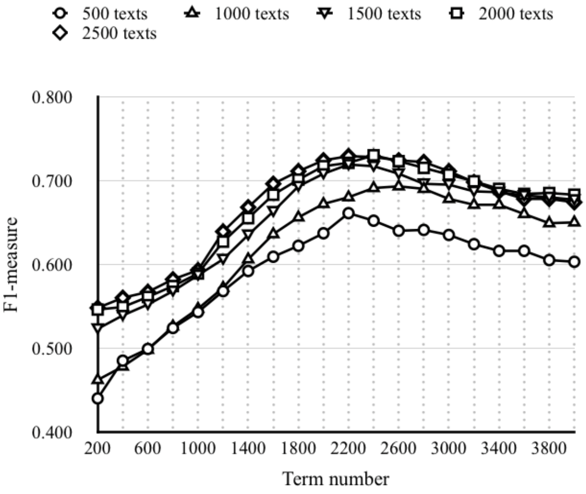

Fig. 1 F1-measure (environment)

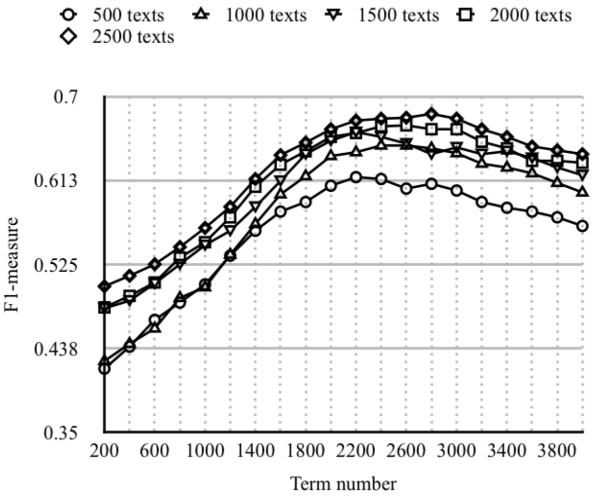

Fig.2 F1-measure (business promotion)

By comparing the F1-measures of two different topics with the same text number (2000), it is obvious that the F1-measure higher of environment topic is higher than that of the business promotion topic. In addition, the term number of which the F1-measures reach their peaks are incompatible but doesn't vary much (environment 2400, business promotion 2600). The results are similar when choosing other text numbers.

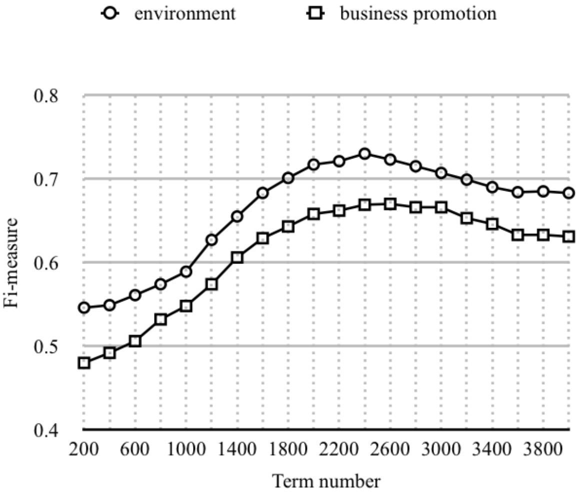

Fig.3 F1-measures comparison

Conclusion

This paper discussed the impacts of quantitative factors on the performance of sentimental classification and the differences of the impacts between two particular topics. As the result demonstrates, the increase of number of texts leads to an incremental tendency of the performance; there is a crest of the number of terms to get to the optimised performance of sentimental classifications. In addition, the prime combination of the two factors for non-identical topics to get a better performance is different as well.

This paper may preliminarily assist those who feel confusing when choosing the particular set of quantifiable factors of the training set when doing research relevant to sentimental classification. Furthermore, it gives evidences and inspirations on how to optimise the performance to analysis the sentiment of textual information from Weibo or some similar social networking forms.

This experiment still remains some defects: since data collecting and manually sentiment marking require huge work in the preparation stage, the topic discussed is limited to only two different ones. In the future, more topics should be studied to further prove to results of this paper and be subdivided to discuss the optimal method to implement sentimental classification of better performance and accuracy. In addition, only CHI-square test is used to extract the sentimental terms, and there are alternatives such as information gain (IG); the weight of terms is simply represented linearly by document frequency (DF), and other possible solution like TF-IDF could be discussed; SVM is not the only way to be applied to sentimental classification, other methods could also be studied to improve the performance of textual analysis.

References

[1] Liu Z. Empirical study of sentiment classification for Chinese microblog based on machine learning[J]. Computer Engineering & Applications, 2012.

[2] Zhang Q, Zhang L, Dong S. Effects of category distribution in a training set on text categorization[J]. Journal of Tsinghua University, 2005.

[3] Wu D. A search algorithm for SVM training sample dataset[J]. Computer Applications & Software, 2010.

[4] Li X, Cao H, Huang L. Study about effect of relevant quantitative indexes of training set in text classification[J]. Application Research of Computers, 2014.