Research on sentiment classification of micro-blog short text based on topic clustering

Shifen Tu1, Bo Yang1∗
1 School of Information Management, Jiangxi University of Finance and Economics, Nanchang City, Jiangxi Province, 330029, China
∗Corresponding author’s e-mail: yangbo@jxufe.edu.cn

Abstract. Aiming at the shortage of research on micro-blog short text fine-grained sentiment classification, a fine-grained sentiment classification method about micro-blog short text based on PLSA model and K-means clustering model was proposed. PLSA is used to calculate the probability matrix between documents and topics, words and topics in the corpus. In terms of the probability distribution of words and topics, K-means algorithm is used to cluster the probability distribution of words on topics and merge the similar topics. Based on the sentiment ontology library, emotion recognition is carried out for the merged topics. Then, according to the merged document and topic probability matrix, the document sentiment category is classified. The experimental results show that the sentiment analysis method integrated with PLSA and K-means can obtain higher classification accuracy than the PLSA model method alone.

1. Introduction
In the Internet era, social media such as online news comments and microblogs have become tools and platforms for people to share their views and experiences. Microblog has become the main Internet social media by making use of its huge user base and strong communication ability. Network users post short texts on microblog to express their opinions and feelings about popular events. Sentiment analysis of these comments can be applied to public opinion management, public opinion poll, commercial marketing intelligence and other fields, and has broad application space and development prospects. It is of great significance to study the emotion recognition and automatic classification of micro-blog short text by using computer technology to realize the emotion recognition of massive microblogs.

The existing research methods of short text sentiment classification mainly include methods based on knowledge base, machine learning, topic model. Chen Ke[1] proposed a self-training semi-supervised sentiment classification method based on multiple classifiers combination. They used the weight of sentiment contribution of the classifier as the confidence and samples with high confidence as the training set to train and obtain a better classifier, and experimented on the micro-blog short texts corpus. Yang Aimin proposed an automatic sentiment classifier for micro-blog short texts based on keywords and probability calculations, and adopted an algorithm combining method integration and voting[2]. Xu Jun[3] made improvements in feature engineering based on the characteristics of the short texts of web news comments, selected different feature sets, multiple feature dimensions, feature weight calculation methods and part of speech, integrated these feature factors, and applied them to automatic sentiment classification. Yang Jianneng[4] made full use of semantic computing of text to automatically annotate sentiment categories of micro-blog short text. In terms of using emotional knowledge base, Moreo A[5] used sentiment lexicon to analyze the emotional tendency of netizens in multiple dimensions. Penalver-Martinez[6] put forward the feature-based method of opinion mining. Pang and Lee designed a variety
of classifiers, such as NB, ME and SVM, in the short text corpus of movie reviews and achieved better classification results under different feature extraction algorithms\cite{7}. Ni chose the CHI method and the information gain method to extract the features of the text, and then designed the text sentiment classifier based on the NB, SVM and Rocchio algorithm\cite{8}. Thet used the syntactic structure to convert long sentences into clauses, judged the emotional memory of the clauses through a rule-based algorithm, and then calculate clauses' emotional memory value. Finally, a classifier was designed to achieve a better effect of emotion classification in short text of movie reviews\cite{9}.

This paper takes seven kinds of short-text emotion, namely, happiness, good, anger, sorrow, fear, evil, and shock as emotion categories, and proposes a micro-blog short text sentiment classification method based on topic clustering. The topic model probabilistic latent semantic analysis (PLSA) and K-means clustering method are combined to classify short texts. The document-topic and word-topic probability matrices of corpus set are calculated by PLSA. Based on the distribution of words on the topic, K-means clustering is carried out for words, and then merge similar topics, determine the emotional category of the topic based on the sentiment ontology library, and use the document-topic probability distribution as the benchmark to realize the sentiment classification of short texts. At the same time, two methods of topic clustering and topic model only are analyzed and the experimental conclusions are drawn.

2. Process of short text sentiment classification based on topic clustering

The sentiment classification process of micro-blog short texts based on topic clustering is shown in Figure 1. Firstly, make pretreatment on the micro-blog short text set obtained by crawler, filter out stop words and useless words after word segmentation, and perform word frequency statistics to obtain the document-word matrix. Then, use the PLSA model to calculate the probability matrices of "word-topic" and "document-topic". Based on the K-means algorithm and the "word-topic" probability distribution, clustering the probability distribution of the words on the topic. Then define the topic with more co-occurring emotional words in the same cluster as for the same sentiment topic, combined with the sentiment ontology library. The topics are merged to generate new probability matrices of "word-topic" and "document-topic". According to the point of view of literature\cite{11}, probability distribution is a direct correlation between topic and words, and there are many words of the same kind of emotion under the same topic. Therefore, for the merged "word-topic", the emotion category of the topic can be judged based on the sentiment ontology library. And because the document with the highest probability under a certain topic has the same sentiment as the topic, after the topic sentiment is determined, the automatic classification of the short text sentiment category can be realized according to the probability distribution of the "document-topic".
3. Probability matrix extraction based on PLSA

The PLSA model can be used to extract two probability matrices of "document-topic" and "word-topic" from a large amount of text [12]. Firstly, given document set $D=\{d_1,d_2,\ldots,d_n\}$ and word set $W=\{w_1,w_2,\ldots,w_n\}$, $freq=\{d_i,w_j\}$ represents the occurrence probability of word $w_j$ in document $d_i$, then the "document-word" co-occurrence matrix $M_{d,w}=[freq(d_i,w_j)]$. Suppose the topic category is $Z=\{z_1,z_2,\ldots,z_k\}$, and $k$ is the number of topics. The PLSA model assumes that the probabilities between words and documents, topics and documents or between words are subject to conditional independence, and the corresponding joint distribution probability is

$$p(d_i,z_k,w_j) = p(d_i)p(z_k \mid d_i)p(w_j \mid z_k)$$  \hspace{1cm} (1)

$p(d_i)$ represents the probability of selecting document $d_i$, $p(z_k \mid d_i)$ represents the probability of a certain topic $z_k$ appearing in a given document $d_i$. $p(w_j \mid z_k)$ represents the probability of the occurrence of word $w_j$ under the condition of given topic $z_k$. In this paper, event $Ev_t$ is obtained based on the probability distribution of "word-topic". According to Bayes rule, we can get:

$$p(d_i,w_j) = p(d_i)\sum_{k=1}^{k} p(z_k \mid d_i)p(w_j \mid z_k)$$  \hspace{1cm} (2)

Fit the latent semantic model using the Expectation Maximization (EM) algorithm. After initialization with random number, step E and step M are executed alternately for iterative calculation. The prior probability of latent semantic $Z$ generated by step E:

$$p(z_k \mid d_i,w_j) = \frac{p(z_k \mid d_i)p(w_j \mid z_k)}{\sum_{i=1}^{i} p(z_k \mid d_i)p(w_j \mid z_k)}$$  \hspace{1cm} (3)

In the step M, recompute $p(w \mid z)$ and $p(z \mid d)$ matrix according to the $p(z \mid d,w)$:
The logarithm of the likelihood function is as follows:

\[
L = \sum_{j=1}^{M} \sum_{j=1}^{N} \text{freq}(d_j, w_j) \log(p(d_j, w_j))
\]

When the increment of the expected value of the likelihood function \( L \) is less than the threshold, the iteration is terminated. So we have an optimal solution:

\[
p(w | z) = \left[ p(w_j | z_k) \right] \text{max} \quad \text{and} \quad p(z | d) = \left[ p(z_k | d_i) \right]
\]

### 4. Experiment and result analysis

#### 4.1 Experimental data acquisition

The experiment crawled Sina microblog comments about "Pan Shiyi's donation to Harvard", and obtained 1964 microblog comments. In this paper, the sentiment ontology library is used as the sentiment knowledge base for sentiment classification, so in the process of corpus processing, comments without emotion words and text length less than 10 words are directly removed. We get 1521 corpus, and 1500 of them are randomly selected for marking. Five relevant professionals were invited to manually annotate the corpus. The content is marked as 7 types of emotions. The annotation results are voted. If 3 people have the same annotation results, the corpus annotation is considered valid. Finally, 1300 microblog comments are used as the experimental corpus in this paper, the distribution of the corpus is shown in Table 1.

| emotion | happiness | good | anger | sorrow | fear | evil | shock |
|---------|-----------|------|-------|--------|------|------|-------|
| Document number | 98 | 121 | 425 | 159 | 92 | 320 | 85 |

#### 4.2 Determination of the clustering number and topics

The number of topics in this paper is set to increase from 7 to 28. The number of clusters also started from 7 to 28 to explore the optimal parameter combination. The classification accuracy is used as the evaluation standard, and the calculation formula is shown in equation (7).

\[
\text{Accuracy} = \frac{\text{Correctly classified comments}}{\text{Total number of comments}}
\]

According to equation (7), in the corpus selected in this paper, the accuracy of text sentiment classification is calculated as the number of topics and clusters increase, and the finally result is the accuracy of "topic number-cluster number" matrix. To facilitate the display, here we only select the number of clusters with the highest classification accuracy under a certain number of topics. As shown in Figure 2, the content of the x-axis is (number of clusters, number of topics), which represents the highest accuracy rate that can be obtained under a certain number of topics and the number of clusters of the K-means algorithm, and the y-axis represents the accuracy of sentiment classification.
4.3 Sentiment classification results of comment text based on topic clustering
Set the number of clusters to 12 and the number of topics to 19. In the contrast experiment, the number of topics was 17. Two text emotion classification methods, PLSA&K-means and PLSA, are used respectively. The accuracy of corpus classification under various emotions is shown in Table 2.

| emotion     | happiness | good   | anger  | sorrow | fear   | evil   | shock  | average | total  |
|-------------|-----------|--------|--------|--------|--------|--------|--------|---------|--------|
| PLSA+K-means| 91.67     | 96.88  | 93.55  | 93.79  | 90.77  | 96.64  | 94.20  | 93.93   | 95.23  |
| PLSA        | 84.52     | 91.37  | 87.10  | 88.28  | 84.62  | 89.60  | 89.13  | 87.80   | 89.17  |

From the experimental results in Table 2, it can be seen that PLSA+K-means has significantly higher accuracy than the PLSA algorithm. Based on the method of this article, the integration of topic word extraction technology can deeply study the emotional distribution of netizens under different topics, which is of great significance to the analysis of online public opinion.

5. Conclusions
This paper proposes a short text sentiment classification method based on topic clustering. It uses a technical solution combining topic model PLSA and clustering algorithm K-means to cluster words in the "word-topic" probability matrix and merge similar topics. The emotion categories of the combined topics were determined by using the emotion ontology database as the support of emotion knowledge, and the fine-grained sentiment classification of short texts was realized by using the probability matrix of "document - topic". The experimental results show that PLSA&K-means has better classification effect than PLSA. Based on the method in this paper, which integrates the keyword extraction, can be applied to key technologies of public opinion such as subtopic detection, sentiment classification and sentiment strength detection. At the same time, it can be applied to the automatic annotation of sentiment categories in large-scale short texts, which is of great significance to the construction of fine-grained sentiment corpus.

Acknowledgment
This work is supported by the innovation and entrepreneurship training program for college students in Jiangxi Province(202110421115),the Key Projects of Jiangxi Province Social Science Planning(19TQ01).

References
[1] CHEN Ke, LI Shujun, XIE Bo. Sentiment Analysis of Chinese Micro-blog Based on Semi-Supervised[J]. Computer & Digital Engineering, 2018, 46(09): 1850-1855.
[2] YANG Aimin, ZHOU Yongmei, ZHOU Jianfeng. Automatic annotation methods for Chinese micro-blog corpus with sentiment class[J]. Journal of Computer Applications, 2014, 34(8):2188-2191.

[3] XU Jun, DING Yuxin, WANG Xiaolong. Sentiment Classification for Chinese News Using Machine Learning Methods[J]. Journal of Chinese Information Processing, 2007, 21(6):1011-1014.

[4] YANG Jianeng, YANG Aimin, ZHOU Yongmei. Sentiment classification method of Chinese Micro-blog based on semantic analysis [J]. Journal of Shandong University (Natural Science), 2014(11):1671-9352.

[5] Moreo A, Romero M, Castro JL, et al. Lexicon-based Comments-oriented News Sentiment Analyzer system[J]. Expert Systems With Applications, 2012, 39(10):9166-9180.

[6] Penalver-Martinez I, Garcia-Sanchez F, Valencia-Garcia R, et al. Feature-based opinion mining through ontologies[J]. Expert Systems With Applications, 2014, 41(13):5995-6008.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]// Acl-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2002:79-86.

[8] Ni X, Xue G R, Ling X, et al. Exploring in the weblog space by detecting informative and affective articles[C]// International World Wide Web Conference, 2007:281-290.

[9] Thet T T, Na J C, Khoo C S G. Aspect-based sentiment analysis of movie reviews on discussion boards[J]. Journal of Information Science, 2010, 36(6):823-848.

[10] XU Linhong, LIN Hongfei, PAN Yu, et al. Constructing the Affective Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2):180-185.

[11] LIN Jianghao, GU Yeli, ZHOU Yongmei, et al. Automatic annotation of emotions based on short texts from news comments[J]. Computer Systems & Applications, 2019, 28(01):207-211.