ORIGINAL RESEARCH PAPER
Volume-7 | Issue-11 | November-2018 | PRINT ISSN No 2277 - 8179
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

UNSUPERVISED SENTIMENT ANALYSIS ON CHINESE MICROBLOG BASED ON TOPIC SENTIMENT MODEL

Computer Science

Runyu Liang School Of Computer Science And Software Engineering, Tianjin Polytechnic University, Tianjin, 300387, China

Lanhe Yang* School Of Computer Science And Software Engineering, Tianjin Polytechnic University, Tianjin, 300387, China *Corresponding Author

ABSTRACT

Sentiment analysis in microblog has important theoretical and application value in personalized recommendation and public opinion analysis.In order to solve the problem of feature sparsity of microblog corpus, Then the paper uses the weight of the words in corpus as one of the BTSM's (Biterm Topic-Sentiment Model) parameter to form the unsupervised W-BTSM model finally. The model adds a sentiment layer to Biterm Topic Model and fuses the weighted model with it ,thus a three-layer Bayesian model of "sentiment-topic-term" is formed. It directly models the generation process of biterm in microblog corpus. The experimental results on the NLP & CC2012 corpus and the real micro-blog data obtained through web crawler show that W-BTSM model can effectively identify the sentiment tendency of Chinese microblog, the F-measure of W-BTSM model is higher than ASUM model and JST model.

KEYWORDS

Sentiment analysis; microblog;feature sparsity;unsupervised;W-BTSM

INTRODUCTION

With the rise of Web 2.0 and the popularity of the Internet, the Internet social services have also been developed rapidly, microblog is favored by the majority of Internet users. A large number of microblog messages are published every day. Through these microblog comments, microblog users, enterprises and governments can quickly understand people's sentiment attitudes toward the topic. Sentiment classification of microblog messages will play a great role in our daily life and even the whole society.

With the explosive growth of microblog comments, we can no longer rely solely on manual intervention to deal with the huge text data of microblog. The text of comments submitted by Internet users has become more and more short and random.The grammatical structure is not standardized. It has strong sparsity and incompleteness, making the extraction of valuable information become more and more difficult.According to these characteristics of Weibo comments, the main content of this paper is the following three points. And figure 1 is the flow chart of sentiment classification for Chinese microblog.

1. This paper proposes an unsupervised W-BTSM model combining weight model with biterm topic-sentiment model. In this way, the influence of short text and stopwords has been eliminated.

2. The parameters are estimated by the Gibbs sampling method. Obtaining the sentiment polarity of documents by solving the model.

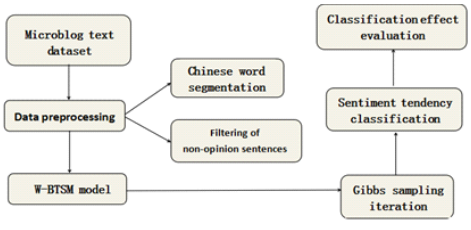

Figure 1: flow chart of sentiment classification for Chinese microblog

3. We demonstrate the effectiveness of the W-BTSM model in sentiment classification by experimenting on the NLP & CC2012 corpus and the real micro-blog data sets.

RELATED WORK

In 2008, Pang and Li et al.[1] pointed out that identifying themes plays an important role in sentiment classification. The sentiment generation model considering the topic has begun to attract the attention of researchers. In 2010, Lin et al.[2] proposed an LDA-based Joint Sentiment-Topic Model (JST). The model adds an emotional layer to the LDA model and expands into a four-layer Bayesian network. Jo[3] proposed the ASUM (Aspect and Sentiment Unification Model) model, which is similar to JST, but it assumes that the Sentiment polarity of the words in each sentence is the same. The words in the same sentence are from the same language model. However, JST and ASUM mainly consider the topic distribution of a single document . They are more suitable for sentiment classification of long text.

In 2013, Yan et al.[4] proposed a topical model for short texts, the Biterm Topic Model (BTM). The topic learning process of the model does not need any external data. For the short text topic model, it directly models the double words in the corpus. The impact of this feature sparse on topic modeling. Subsequently, Xie Jun et al.[5] put forward a joint and sentiment model BJSTM (Biterm Joint Sentiment Topic Model) for short text. The model can effectively extract the keywords of each sentiment, and get a better classification accuracy of sentiment.

FILTERING OF NON-OPINON SENTENCES

Before processing the acquired data, this paper observes and analyzes a large number of real comments of hot topics on microblog, and finds that some content is obviously not with viewpoints containing emotional tendencies, such as: statement of objective facts, sales advertising, etc. Filtering out these meaningless microblog content can improve the efficiency and results of classification processing before classification. For this purpose, this article has developed the following filtering rules:

1. There is no text content in the sentence, or only punctuation, pictures or links are meaningless microblogs;

2. Comments containing "data display", "investigation indication", and special punctuation "[", "]" are judged as non-view sentences.Such microblogs are usually of introductory nature, or the probability of opinion sentences is extremely low. This article directly filters it.

3. Almost all the micro-blog news that contains words such as special offer, discount and sale is advertising. Filter out these sentences.

4. “@username” only serves as a reminder and emphasis, and does not reflect any hot topics or emotional information. For this purpose, all “@usernames” are removed and other texts are retained.

5. There are a lot of forwarding behaviors in Sina microblog. "//@username:" means that the current Weibo is a microblog message that forwards others. This part of Weibo news does not make any sense to our sentiment analysis. This article removes all the content of the forwarding Weibo part, leaving only the most original Weibo message.

WEIGHT MODEL

A stop word is a word that appears frequently in the text but does not contain any emotional information. For the stop words in the Weibo comment dataset, the traditional approach is to use the stop word table matching removal. Using different stopword tables to process results differently, this paper uses an improved weight model instead of the traditional stopword matching method. The difference coefficient is used to process the vocabulary in the document, making the selection of stopwords more scientific, and also making up for the lack of word frequency and document frequency.

The coefficient of variation is expressed as the ratio of the standard deviation to the mean. The specific calculation method is as shown in formula (1), which describes the degree of dispersion of document frequency in which each word appears in different categories.

$$V(\omega)=\frac{\sqrt{\sum_{m=1}^{M}(f_{d_m}-\bar{f})^2/n}}{\bar{f}} \quad (1)$$

In formula 1, f_{d_m} represents the number of occurrences of the word in the document , $\bar{f}$ represents the average of the number of occurrences of the word in each document, and n represents the number of documents in which the word ω appears in the document. When the coefficient of variation is larger, the greater the degree of dispersion of the word, the greater the weight, indicating that the word should not be used as a stopword to a large extent; otherwise, the smaller the difference coefficient, the smaller the degree of dispersion between words and the smaller the weight. Words do not contribute to the emotional classification of documents and should be considered as stop words.When the difference coefficient is smaller, the weight is also smaller. The word does not contribute to the emotional classification of the document and should be regarded as a stop word.

W-BTSM MODEL

The general topic sentiment model is to model the process of generating a single document. The model used in this paper fuses the word weight model with the biterm topic sentiment model to form a three-layer Bayesian unsupervised topic sentiment model of "sentiment-topic-term". It not only effectively solves the sparsity and incompleteness of short texts, but also does not reduce the subject relations between words. The difference coefficient of each word has been known before searching for biterm in the document. At this point, the biterm in the model no only represent the two co-occurrence words in the document, but also indicate whether the words are stowords. For example, a document is made up of words $\omega_1,\omega_2,\omega_3$, and its biterm are changed from $\{(\omega_1,\omega_2),(\omega_1,\omega_3),(\omega_2,\omega_3)\}$ $\{(V(\omega_1),V(\omega_2),),(V(\omega_1),V(\omega_3),),(V(\omega_2),V(\omega_3),)\}$ to. This simplifies the work of text preprocessing, and can be used as both stop word processing and text sentiment classification in the process of model application.Figure 2 is a model diagram of the W-BTSM model. Table1-3 is a symbolic description used in the model.

Figure 2: W-BTSM model

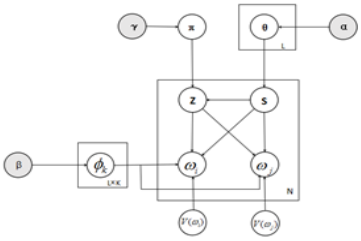

TABLE – 3 symbolic description

symbol	description	symbol	description
M	Number of documents	S	sentiment
N	Number of biterm	π	Sentiment distribution of corpus
K	Number of topics	θ	topic distribution of sentiment
W	Number of words	ϕ	vocabulary distribution of topic
L	Number of sentiment categories	γ	Dirichlet distribution prior parameters of π
$\omega_{i,j}\,(j\subset[1,2])$	word	α	Dirichlet distribution prior parameters of θ
Z	topic	β	Dirichlet distribution prior parameters of θ

Assuming that there are M documents in corpus $D=\{d_1,d_2,\ldots,d_N\}$

The biterm set $B=\{b_1,b_2,\ldots,b_N\}$ is extracted from D ($b_i=\{\omega_{i,1},\omega_{i,2}\}$).

There are K themes $Z=\{z_1,z_2,\ldots,z_K\}$ and L emotions $S=\{s_1,s_2,\ldots,s_L\}$.

The process of generating documents in the BJSTM model is as follows.

1. For the entire corpus D:
Draw $\pi\sim$ Dirichlet(γ)

2. For each sentiment S:
Draw $\theta_s\sim$ Dirichlet(α)

3. For each topic Z:
For each sentiment S:
Draw $\phi_{z,s}\sim$ Dirichlet(β)

4. For each word:
Calculate the weight of the word

5. For each biterm:
(1) Draw $s_i\sim$ Multionmial $(\pi)\cdot V(\omega)$

(2) Draw $z_i\sim$ Multionmial $(\theta)\cdot V(\omega)$

(3) Draw $\omega_{i,1}$, $\omega_{i,2}\sim$ Multionmial $(\phi)\cdot V(\omega)$

Through Gibbs sampling iterations, the parameters a, b, and c are estimated as equations 2 through 4. Where $n_{l,k,\omega,\neg i}$ represents the number of words w in the set of words whose emotion is l and the topict is k, except for the word that the number is I.

$$\pi_l=\frac{n_{l,\neg i}+\gamma}{N_{\neg i}+L\gamma} \quad (2)$$

$$\theta_{l,k}=\frac{n_{l,k,\neg i}+\alpha}{n_{l,\neg i}+K\alpha} \quad (3)$$

$$\varphi_{l,k,\omega}=\frac{n_{l,k,\omega,\neg i}+\beta}{n_{l,k,\neg i}+W\beta} \quad (4)$$

In the W-BTSM model, we need to use the biterm's emotion to estimate the emotion of the document. If there are N_d biterm ($b_{i_j}, j\subset[1,N_d]$) in document d, then the ratio of emotion in document d is equation 5.

$$P(s=l/d)=\sum_{j=1}^{N_d}P(l/b_{i_j})P(b_{i_j}/d) \quad (5)$$
$$=\sum_{j=1}^{N_d}\frac{\pi_l\theta_{l,k}\phi_{l,k,\omega_{i_j,1}}V(\omega_{i_j,1})\phi_{l,k,\omega_{i_j,2}}V(\omega_{i_j,2})}{\sum_l\pi_l\theta_{l,k}\phi_{l,k,\omega_{i_j,1}}V(\omega_{i_j,1})\phi_{l,k,\omega_{i_j,2}}V(\omega_{i_j,2})}\cdot\frac{n_d^{(b_{i_j})}}{N_d}$$

The emotional tendency of the document is the emotion with the highest proportion of emotions in document d.

$$S_d=\arg\max(P(s/d)) \quad (6)$$

EXPERIMENTS

The experiment was implemented in a Python development environment. Use jieba segmentation to perform Chinese word segmentation on the corpus. The specific settings of the W-BTSM model parameters are as follows: $\alpha=2$, $\beta=0.01$, $\gamma=0.1$. The number of iterations is 1000. According to the actual situation of the corpus, the number of topics K is set to 25. The parameters in the comparison model JST and ASUM are set according to the original literature method. The number of iterations is 1000.

The experimental data was selected from NCP&CC2012 test corpus and microblog data obtained through web crawler, and the test data was divided into three groups.Each set of data contains 3,000 positive microblogs and 3,000 negative microblogs.The evaluation index is the precision, recall and F-measure provided by the NLP&CC2012 evaluation.

$$\text{Precision}=\frac{\text{The number of correctly classified documents}}{\text{Total number of documents classified}} \quad (7)$$

$$\text{Recall}=\frac{\text{The number of correctly classified documents}}{\text{Total number of documents manually labeled}} \quad (8)$$

$$\text{F-measure}=\frac{2\times \text{Pr}\,ecision\times \text{Re}\,call}{\text{Pr}\,ecision+\text{Re}\,call} \quad (9)$$

Three models of W-BTSM, JST and ASUM were tested on the above three groups of corpora.The final comparison results are shown in Tables 2, 3, and 4.

TABLE –2 Sentiment classification precisions comparison

Precision	Data1	Data2	Data3	P_{avg}
JST	0.5776	0.5251	0.5563	0.5530
ASUM	0.7285	0.7075	0.7682	0.7347
W-BTSM	0.7633	0.8102	0.7789	0.7841

TABLE – 3 Sentiment classification recalls comparison

Recall	Data1	Data2	Data3	R_{avg}
JST	0.7239	0.6854	0.7136	0.7076
ASUM	0.7934	0.7241	0.8035	0.7736
W-BTSM	0.8071	0.7893	0.7564	0.7842

TABLE – 4 Sentiment classification F-measures comparison

F-measures	Data1	Data2	Data3	F_{avg}
JST	0.6425	0.5946	0.6252	0.6208
ASUM	0.7590	0.7157	0.7855	0.7534
W-BTSM	0.7846	0.8047	0.7675	0.7856

From the table we can see that the performance of these three models on different data sets is slightly different. The average value of F-measures in the W-BTSM model is 78.56%, which is about 3% higher than the ASUM model and 16% higher than the JST model. In general, the W-BTSM model has higher index values than the other two models. From the experiments above,it is evident that the method proposed in this paper is effective forsentiment classification of Chinese microblog.

CONCLUSIONS AND FUTURE WORKS

The method of traditional unsupervised emotion tendency classification can not solve the sparsity of microblog short texts well. This paper proposes to use the unsupervised model W-BTSM to complete emotion tendency classification of Chinese microblog. By constructing the co-occurrence relationship of vocabulary in the corpus and using the method of joint topic sentiment recognition, the matrix sparsity and imperfection of short text sentiment classification are better solved. Through the real datasets, it is proved that the proposed method has a good effect on emotion classification of Chinese microblog. However, the model used in this paper does not consider the contextual semantic information. In the emotion recognition, some information may be lost and affect the classification result. Therefore, this should be used as the next research direction to improve the classification performance.

REFERENCES:

[1] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.
[2] Chenghua Lin, Yulan He, Richard Everson. A comparative study of bayesian models for unsupervised sentiment[C]. Proceedings of the Fourteenth Conference on Computational Natural Language Learning, 2010: 144-152.
[3] Yohan Jo, et al. Aspect and Sentiment Unification Model for Online Review Analysis[C], WSDM, Hong Kong, China, 2011.
[4] Yan X, Guo J, Lan Y, et al. A Biterm Topic Model for Short Texts[C]. Proceedings of International Conference on World Wide Web. New York, NY,USA: ACM, 2013. 1445−1456
[5] XIE Jun, HAO Jie, SU Jingqiong, et al. A Joint Topic and Sentimen Model for Short Texts[J]. Journal of chinese information processing, 2017,31(01):162-168.