Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

An Automatic Construction Approach for Sentiment Dictionary Based on Weibo Emoticons

Xiaohong Hao, Yifan Jia [a)], Qun Gu

School of Computer and Community, Lanzhou University of Technology, Lanzhou 730050, China

[a)]Corresponding author: 958648453@qq.com

Abstract. Different from common texts, microblog texts have large amount of emotions and net-words. Its words and sentences expression are more colloquial and network popular, so traditional sentiment dictionary does not suit for the context of modern microblog short texts. This article puts forward an approach to automatically construct the sentiment dictionary based on microblog emoticons. Collected microblog texts are annotated by emoticons and form sentiment text corpus. To conduct consolidation according to existing sentiment dictionary, extract the sentiment words in the microblog texts per rule of part of speech, calculate the information added value of sentiment words in microblog texts as feature weight, and classify the sentiment words in the method of SVM to get the sentiment dictionary. This article improves the construction method of existing sentiment dictionary. The experimental result shows that the accuracy rate of sentiment dictionary after improvement reaches above 90%, and the overall F value reaches 85%, which are obviously better than existing dictionaries.

Key words: Emoticon; sentiment classification; sentiment dictionary; sentiment polarity; natural language processing.

INTRODUCTION

Since the appearance of Sina microblog in 2009, it gains rapid development. Till Dec 2017, its monthly active users reached 400 million. Each day there're average above 100 million microblog texts sent by these users. To conduct analysis study of orientation of these microblog texts is beneficial for microblog monitoring, public opinion discovery and guidance tec [1]. Due to the nonconformance, loud noise and redundancy of data in microblog texts etc., and considering it usually appears in the form of short text, it brings new challenges and opportunities to network public opinion analysis. Therefore, scholars at home and abroad conduct lots of studies, among which approaches such as SVM (support vector machine) and NB (naive bavesian) etc. are the research hotspot in recent years. The study of text sentiment orientation needs large amount of corpus and high quality sentiment dictionary as the support. Therefore, to construct high quality microblog sentiment dictionary has a vital significance to microblog text orientation study.

As a new communicative means, microblog improves the development and widely application of network symbols and language esp. emoticons [2]. The manifestation form of emoticons is succinct and clear, and can vividly displays various individual emotions, and simplifies language and word comprehension and interpretation in communication and interaction, so it is quite popular among the youth[3]. Therefore, as a special network terminology, the emoticon provides a new thinking to construct social network sentiment dictionary.

RELATED WORK

Currently there're two construction approaches for sentiment dictionary: manual annotation and automatic annotation. Manual annotation construction approach means to summarize sentiment words manually and annotate their sentiment polarity and intensity by reading a large number of texts. Though this approach is too time-consuming,

its constructed sentiment dictionary is of relatively high accuracy rate. Currently most of sentiment dictionaries are constructed by manual annotation. Relatively authoritative sentiment dictionaries are sentiment analysis term set provided by SentiWordNet, General Inquirer (GI) and HowNet etc. Though sentiment dictionaries based on manual construction get relatively good universality, in actual practice it's hard to cover sentiment terms from different fields, and the domain adaption it is not that good. Meanwhile, manual sentiment dictionary construction needs lots of manpower and material resources. So the academic circle focuses more on the auto construction of sentiment dictionary [3].

There are mainly two approaches to automatically construct sentiment dictionary: one based on semantic library, another based on corpus. The approach based on semantic library is to construct a sentiment dictionary with relatively strong universality by finding relations between words (such as synonymy, antonym, superordinate and subordinate relationships etc.). Hu [4] and other scholars made use of seed word set which is known to be commendatory or derogatory, searching for semantic relations between words in semantic library to extend, and get a general sentiment dictionary. Baccianella [5] and other scholars adopted interpretive extension method, and took the paraphrase of synonyms as training corpus to study the relation between words in paraphrase. There're lots of approaches to construct sentiment dictionary based on corpus construction, among which conjunction construction and word co-occurrence are the most common methods. Huang [6] and others made use of conjunctions to judge the polarity relation between words, and combined the negative forms of words (such as "Y" and "unY") to establish the constraint matrix of sentiment polarity. And then they made use of PMI (Point wise mutual information) to judge the sentiment polarity of words. This approach applies to corpus with relatively subjective sentences and obvious successive emotional variety, but it cannot judge the sentiment polarity for statements with adversative relation. Li Yonggan [7] etc. organized and summarized the results of text dependency based on Chinese dependency syntax analysis, and then made use of some dependency rules to extract sentiment words and make polarity judgment. Turney [8] etc. made use of search engine to retrieve sentiment words, and then calculated PMI value, thus to find out the word with the closest meaning with the known words to constitute a synonym dictionary. Sentiment dictionary constructed in this approach is with relatively high accuracy rate. But its calculation speed is relatively slow. The approach to automatically construct sentiment dictionary based on microblog emoticons belongs to a kind of approach based on corpus. It's mainly to annotate microblog texts or sentiment words by emoticons to calculate the interrelation degree between extracted sentiment words and microblog texts, to get the polar intensity of sentiment words. Ma Bingnan[9]etc. put forward an approach to make use of emoticons to extract text sentiment dictionary, collect sentiment words with cross-media idea, introduce crowdsourcing annotation emoticons, and make use of co-occurring mutual information computing method to extract dictionaries with different emotions. Sentiment dictionary generated in this approach is of much higher accuracy rate and better effect than Tsinghua University dictionary, Hownet dictionary, Dalian University of Technology and Taiwan University, but crowdsourcing annotation needs lots of manual participation. Gui Bin [10] etc. introduced emoticons to annotate the emotion tendency of microblog texts, calculate chi-square statistics of sentiment words to gain the sentiment polarity strength, and generate the sentiment dictionary by judging the tendency of sentiment words according to the probability of occurrence of sentiment words in positive and negative microblog texts. This approach greatly induces human intervention, but more makes use of Chinese semantic rules to affect the accuracy rate of sentiment dictionaries. On this basis, this article changes the judging method of sentiment word orientation, adopts SVM approach to conduct binary classification of sentiment words to improve the accuracy rate.

CONSTRUCTION PROCESS OF SENTIMENT DICTIONARY

The construction approach for sentiment dictionary in this article mainly includes 4 parts: data filtering, emoticon extraction and microblog text annotation, sentiment word extraction, and sentiment word dictionary calculation. The main process is as shown in figure 1.

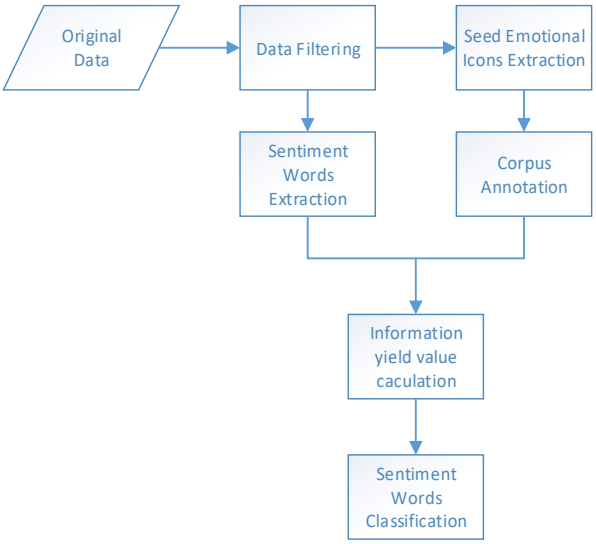

FIGURE 1. Sentiment Dictionary Construction process

This article adopts 4 months' short texts in Sina microblog during Nov 2016 to Feb 2017. It connects to the public data flow API in Sina microblog by web crawler procedure, and extracts the recent data texts in Sina microblog. It then preprocesses them by systematic information filtering, removing duplicate texts, distinguishing words and de-noising etc. to calculate and extract microblog short texts with expressions as the experimental data for this article.

Selection and Processing Rule of Seed Emoticons

Emoticons in social network are words taking pics as the carrier. In modern network language, the meaning of emoticons generally are not same to their Chinese meanings in emoticons. As to emoticons, people mostly intuitively interpret the representative affective feature of these symbols to understand and use these emoticons, but not the Chinese meanings behind these emoticons. For example, the meaning of this emoticon is contempt. This is a negative word in Chinese, but in actual use it's used to express a positive emotion. As to this phenomenon, we need to process these emoticons per the method in the pic, and manually annotate their meanings. To guarantee the accuracy rate of annotation result and reduce the manual handling time, the selection of seed emotions in this article adopts the method of combination of crowdsourcing [11] and document frequency. Firstly it calculates the document frequency including a certain emoticon from microblog corpus, then remove this emoticon from the characteristic space when the DF value of this emoticon is less than a threshold value according to the set threshold. And this article also removes this emoticon from the characteristic space. Then conduct manual annotation in the method of crowdsourcing for selected emoticons, and annotate the sentiment tendency of emoticon to 2 categories: positive and negative.

TABLE 1. Seed Emoticon

Negative Emoticons	Negative Emoticons
	

There're 36 positive emoticons and 15 negative emotions.

Annotated corpus

One microblog corpus may contain both positive and negative emotional tendencies at the same time, so the following rules are put forward according to the occurrence frequency of seed emoticons in the corpus.

Rule 1: If the emotional tendency of seed emoticons in the corpus just includes positive or negative tendency, the emotional tendency of the corpus is decided by this kind of seed emoticons.

Rule 2: If both the positive and negative seed emoticons are included at the same time in the corpus, its emotional tendency is to be decided by the classification of emoticon with the highest frequency of occurrence.

This article selects 2,000 pieces of microblog texts including emoticons in diagram 1. It adopts the rule 1 and 2 to annotate the microblog texts, and makes use of the classification method of voting to determine the final emotional tendency of texts. The final result shows that the results of manual annotation and emoticon annotation are 97% consistent. Thus it's reliable to make use of emoticons to annotate emotional tendencies. This articles uses emoticons to annotate about 100,000 microblog texts, and forms a basic sentiment corpus for study.

Sentiment Word Extraction

Short texts in microblog and traditional texts have quite a big difference. Microblog short texts have lots of network terminologies, which basically don't appear in traditional sentiment dictionaries. Network terminologies are an important component in microblog texts tendencies analysis. Network sentiment words can be extracted per rules below:

Use successive numbers to represent sentences to respond, such as to use "666…" to praise somebody for doing a good job or to resonate to the same online public sentiment.

Use typos to stress lexical semantics. For example, "酱紫" (pronounced "jiangzi" in Chinese) represents "这样子" ("Zhe Yang zi").

Use euphonic numbers or alphabets, such as "1314" ("yi san yi shi" in Chinese pronunciation)to represent "一生一世" ("yi sheng yi shi" in Chinese pronunciation which means forever), "521" (pronounced as wu er yi) to mean "我爱你" ("wo ai ni", which means I love you).

Use abbreviations of pinyin to conceal uncivilized words, such as "BT" to mean "变态" (bian tai, which means abnormal)

Besides, this article organizes sentiment words in traditional dictionaries. It screens vocabularies which are in 3 lexicons at the same time from HowNet, a Chinese sentiment dictionary, NTUSD issued by Taiwan, sentiment word table issued by Dalian University of Technology, and Commendatory Terms Dictionary and Derogatory Terms Dictionary, which means it intersects the above 3 sentiment dictionaries and then finds out and calculates the 3 intersections. There're 3,000 sentiment words in total including collected network sentiment words, which form a basic sentiment dictionary.

Annotation of Sentiment Words

Information gain (IG) method is used to measure the importance of the feature item by quantity of information provided by a feature to the overall classification, then to decide acceptance or rejection of this feature item. Information gain of a feature item refers to the difference of information quantity provided for the overall classification when this feature item exists or does not exists. The information quantity is measured by entropy. Therefore, information gain refers to the value difference between document entropy when considering any feature and that after considering the feature. Information gain in the article reflects the importance of sentiment words in texts with categories, so information gain value is used to calculate the polarity strength of sentiment words. Whether sentiment word is to be added into sentiment dictionary or not is decided by intensity value. For any sentiment word w, use formula (1) to calculate its information gain in positive text, i.e. the strength value of sentiment word polarity.

$$\begin{aligned} Gain(w_i) &= Entropy(S) - Expected\ Entropy(S_{w_i}) \\ &= \left\{-\sum_{j=1}^{M} P(C_j)\right\} - \left\{P(w_i)\times\left[-\sum_{j=1}^{M} P(C_j|w_i)\times\log P(C_j|w_i)\right]\right\} \\ &+ P(\overline{w_i})\times\left[-\sum_{j=1}^{M} P(C_j|\overline{w_i})\times\log P(C_j|\overline{w_i})\right] \end{aligned} \tag{1}$$

In this formula, $P(C_j)$ refers to the occurrence probability of C type document in the corpus. $P(w_i)$ Refers to the probability of document including feature item w_i in the corpus. $P(C_j|w_i)$ Refers to the contingent probability to be type C_j when the document includes feature item w_i. $P(\overline{w_i})$ Refers to the probability of document with no feature item w_i in the corpus. $P(C|\overline{w_i})$ Refers to the contingent probability to be class C_j when document does not include w_i, and M means the number of categories.

It can be known from the definition of information gain that information gain of a feature actually describes the information quantity included to help forecasting categorical attributes. Theoretically speaking, information gain should be the best feature extraction method, but actually there's often low occurrence frequency for many features with relatively high information gains, so when there's little feature items when we search with information gain as the index, it usually has a problem of data sparseness and a relatively poor classification effect. Thus we need to calculate information gain for each word appeared in training corpus, and then appoint a threshold to remove entries with information gain lower than this threshold value from characteristic space.

Information gain value just shows sentiment intensity value, but cannot show the tendency of sentiment words. This article adopts the method of SVM (support vector machine) to classify sentiment tendency for collected sentiment words. First of all, machine learning algorithm is adopted to train the training sample and construct sentiment classifier, then conduct sentiment classification for sentiment words to be recognized by using this sentiment classifier, and extract the information gain value as a sentiment feature weight to get its concept vector of feature to be the input of classifier. Lastly, to get the classification result of sentiment words to be recognized by calculation of classifier as shown in diagram 2.

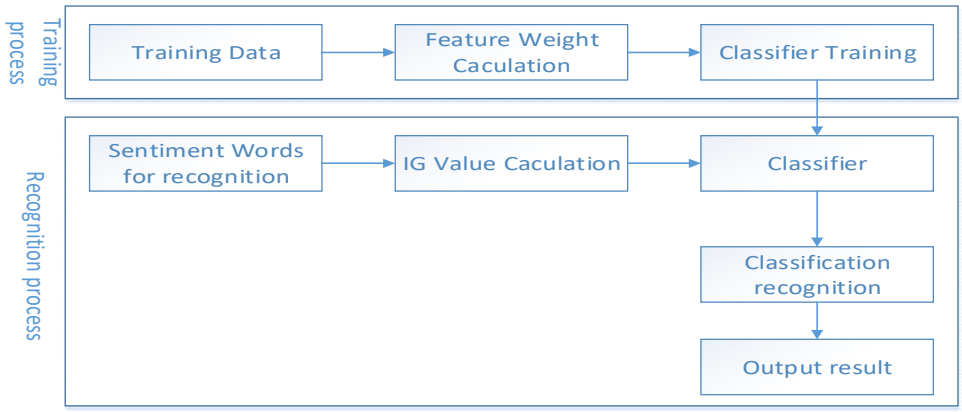

FIGURE 2. Classification Process of Sentiment Words

EXPERIMENTAL RESULT AND ANALYSIS

The experiment goal is to judge the actual classification capacity of the classifier, i.e. the classification accuracy of constructed sentiment dictionary. So P (precision), R (recall) and F (F-measure) values are adopted as test indexes. Basic sentiment dictionary in this article includes 3,000 sentiment words, among which there are 1,800 positive sentiment words and 1,200 negative words. Positive and negative sentiment words shall be judged separately. The accuracy rate and recall rate of positive sentiment words are defined to be:

$$P_{pos} = \frac{N_{posn}}{N_{posnu}} \tag{2}$$

N_{posn}

$$R_{pos} = \frac{N_{posn}}{N_{negn} + N_{posn}} \tag{3}$$

In the formula, refers to the correct annotation returned and the number of positive sentiment words appeared in basic sentiment dictionary. Refers to the number of returned positive sentiment words appeared in positive sentiment word. Refers to the number of negative sentiment word in basic sentiment dictionary returned due to wrong annotation. The calculation of accuracy rate and recall rate of negative sentiment words are same to positive sentiment words, which will not be listed here.

The classifier constructed in the method of SVM divides sentient words into two types: positive and negative, so there'll come out of lots of sentiment words if to directly apply the calculation method in this article. To better analyze the returned sentiment words and meanwhile control the number of sentiment words, threshold value is set to be θ. When the sentiment tendency intensity value of the sentiment word w is larger than θ, w is regarded as a positive sentiment word. When the sentiment tendency intensity of w is less than $-\theta$, w is regarded as a negative sentiment word. When the tendency intensity is between θ and $-\theta$, w is regarded as a neutral word. Experimental result is shown in figure 3 and 4.

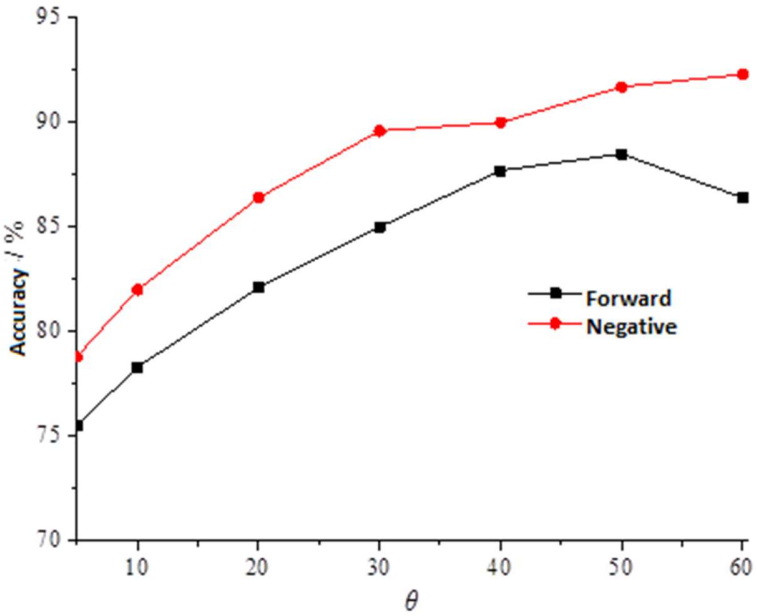

FIGURE 3. Accuracy Rate of Sentiment Dictionary

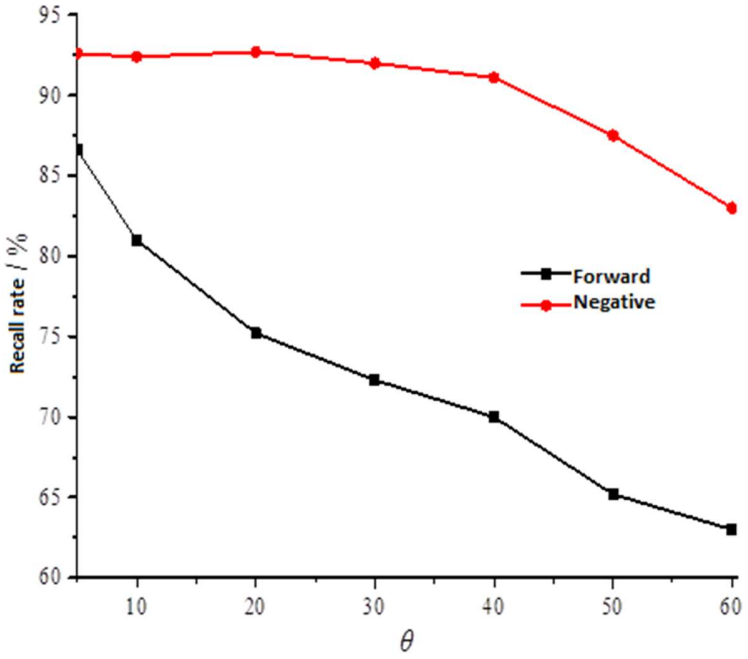

FIGURE 4. Recall Rate of Sentiment Dictionary

It's shown in experimental result that: with the increase of threshold value, the accuracy rate is increasing and recall rate is reducing. It shows that the classification granularity of sentiment words keeps shrinking, and there are more words returned with clear sentiment tendency. In diagram 3, the highest accuracy rate of negative sentiment word reaches 95%, which shows when there's a large microblog corpus, the error in auto annotation exerts a little influence on accuracy rate. The accuracy and recall rate of negative sentiment words are higher than that of positive sentiment words, which is due to quite a lot of positive emoticons extracted and positive sentiment words with auto annotation. Besides, with the increase of threshold value θ, the accuracy rate of sentiment words appears to be in incremental rise, and recall rate is in descending decrease. Generally speaking, accuracy rate and recall rate are a pair of contradictory physical quantities. To improve accuracy rate always needs to sacrifice a certain recall rate, vice versa. Apparently with the increase of threshold value θ, eligible sentiment words become less. And the relatively more sentiment words with stronger sentiment polarity intensity brings less recall rate and more accuracy rate. To better compare sentiment dictionary judgement indexes under different threshold values, set F measure value. F value of positive sentiment word is defined to be:

$$F_{pos} = \frac{2 \times P_{pos} \times R_{pos}}{P_{pos} + R_{pos}} \tag{4}$$

F measure value in negative sentiment dictionary is similar to that in positive sentiment dictionary, which will not be listed here. The positive and negative value of sentiment dictionary generated by different θ values are shown in figure 5. Combine the positive and negative F measure values to define the positive overall F value. It's to be:

$$F = \frac{F_{pos} + F_{neg}}{2} \tag{5}$$

Overall F value of sentiment dictionary is shown in figure 6. It can be seen from figure 6 that when θ is 30, the overall F value in sentiment dictionary reaches 85%, nearly the highest.

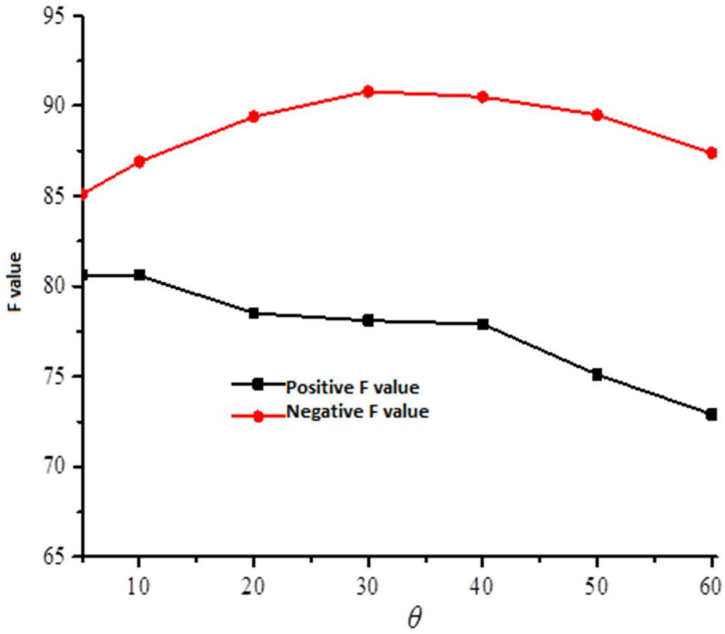

FIGURE 5. Positive and negative F value in sentiment dictionary

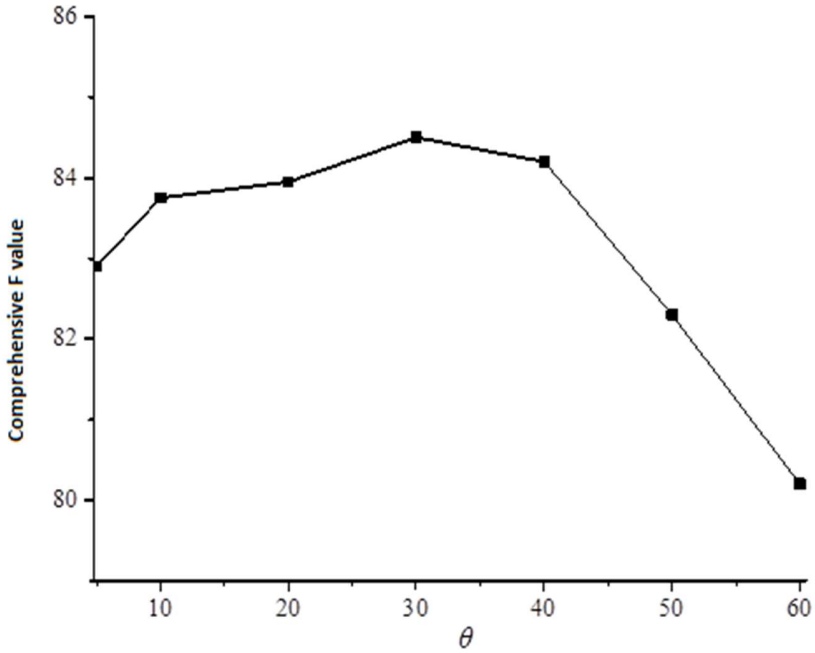

FIGURE 6. Overall F value in sentiment dictionary

CONCLUSION

This article improves current construction approach of sentiment dictionary based on emoticons. In terms of seed emoticons selection, manual annotation and auto annotation are combined to greatly reduce human involvement. In

terms of sentiment dictionary construction, it takes information yield value as the sentiment word polarity intensity value, and improves the accuracy rate of sentiment word polarity annotation. SVM method is used to classify sentiment dictionary, change the original corpus annotation method and sentiment word classification method which excessively depend on Chinese language rules. By improving the generated sentiment dictionary, it gains higher accuracy rate and F value. In the next step, the field feature of microblog texts will be combined to further classify the sentiment dictionary in more details, and construct the field sentiment dictionary of microblog texts.

REFERENCES

1. ZHOU Yong-mei, YANG Ai-min, LIN Jiang-hao. Construction method of Chinese micro-blog emotion dictionary [J]. Journal of shandong university (engineering edition), 2014, 44(3):36-40.
2. WANG Wei, ZHOU Yong-mei, YANG Ai-min, et al. Determination of emotional tendency of weibo emoticons based on seed words [J]. Data acquisition and processing, 2017, 32(1):198-204.
3. TAN Wen-fang. Analysis of the influence of online emoticons [J]. Quest, 2011(10):202-204.
4. Hu M Q, Liu b. Mining and summarizing customer reviews.In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004.
5. Baccianella S, Esuli A, Sebastiani f. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.[C]// International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP, 2010:83-90.
6. Huang S, Niu Z, Shi c. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation [J]. Knowledge-based Systems, 2014, 56(C):191-200.
7. LI Yong-gan, ZHOU Xue-guang, SUN Yan, et al. Method of building emotional lexical database based on dependency analysis and rule statistical analysis [J]. Journal of wuhan university (neo-confucianism), 2013, 59(5):491-498.
8. Turney P d. Thumbs up or Thumbs down? : Semantic orientation applied to unsupervised classification of reviews[C]// Meeting on Association for Computational Linguistics.
9. MA Bing-nan, HUANG Yong-feng, DENG Bei-xing. The social network emotional dictionary structure based on emoticons [J]. Computer engineering and design, 2016, 37(5):1129-1133.
10. GUI Bin, YANG Xiao-ping, ZHANG Zhong-xia, et al. Research on the construction of emotional dictionaries based on weibo emoticons [J]. Journal of Beijing institute of technology, 2014, 34(5):537-541.
11. Ball P. Crowd-sourcing: Strength in Numbers. [J]. Nature, 2014, 506(7489):422.