



Vertical and Sequential Sentiment Analysis of Micro-blog Topic

Shuo Wan¹, Bohan Li^{1,2,3(✉)}, Anman Zhang¹, Kai Wang¹,
and Xue Li^{1,4}

¹ College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
{shuowan, bhli}@nuaa.edu.cn

² Collaborative Innovation Center of Novel Software Technology and
Industrialization, Nanjing, China

³ Jiangsu Easymap Geographic Information Technology Corp., Ltd.,
Nanjing, China

⁴ School of Information Technology and Electrical Engineering,
University of Queensland, Brisbane, Australia

Abstract. Sentiment analysis of micro-blog topic aims to explore people's attitudes towards a topic or event on social networks. Most existing research analyzed the micro-blog sentiment by traditional algorithms such as Naive Bayes and SVM based on the manually labelled data. They do not consider timeliness of data and inwardness of the topics. Meanwhile, few Chinese micro-blog sentiment analysis based on large-scale corpus is investigated. This paper focuses on the analysis of sequential sentiment based on a million-level Chinese micro-blog corpora to mine the features of sequential sentiment precisely. Distant supervised learning method based on micro-blog expressions and sentiment lexicon is proposed and fastText is used to train word vectors and classification model. The timeliness of analysis is guaranteed on the premise of ensuring the accuracy of classifier. The experiment shows that the accuracy of the classifier reaches 92.2%, and the sequential sentiment analysis based on this classifier can accurately reflect the emotional trend of micro-blog topics.

Keywords: Vertical sentiment analysis · fastText · Distant supervision
Sequential analysis

1 Introduction

Sina Weibo, the largest social network platform in China, has more than 150 million daily active users and an average of two hundred million micro-blogs published daily. People can push their life and opinion to Weibo, or comment on hot events. These

This work is supported by National Natural Science Foundation of China (61672284, 41301407), Funding of Security Ability Construction of Civil Aviation Administration of China (AS-SA2 015/21), Fundamental Research Funds for the Central Universities (NJ20160028, NT201 8028, NS2018057).

© Springer Nature Switzerland AG 2018

G. Gan et al. (Eds.): ADMA 2018, LNAI 11323, pp. 353–363, 2018.

https://doi.org/10.1007/978-3-030-05090-0_30

subjective data bring great convenience to the study of sentiment analysis. The real-time sentiment information mining of micro-blog can accurately reflect the trend of micro-blog topics and provide early warnings, which has positive significance for individuals, businesses and governments.

At present, most work about sentiment analysis focus on improving the performance of classification algorithms [1] based on existing datasets, especially the sentiment analysis of Chinese micro-blog, which lacks sound sentiment lexicons and large-scale training corpus. This paper mainly discusses the sequential sentiment analysis based on large-scale Chinese micro-blog corpus. The popular micro-blog topic of “ZTE Denial Order” was selected for detailed sequential analysis, which fully excavated the timeliness of micro-blog data. Main work of this paper are as follows:

- (1) We designed a micro-blog general crawler and collected 35 million Sina Weibo data. A training set and test set for a Chinese micro-blog sentiment analysis was constructed by using a method of distant supervision based on Emoji. Meanwhile, we use the Sentiment Lexicon to further filter the training set.
- (2) fastText [2] is used to train word vector and generate classification models. Experiment shows that based on the classification model trained by the training set, the accuracy reaches 92.2% on the test set.
- (3) We designed a micro-blog specialized crawler, which collects data on a vertical topic on Weibo and performs dynamic sentiment analysis on timeline. The experiments have proved that the sequential sentiment analysis can accurately reflect the micro-blog user’s attitude toward a topic or event and the trend of the incident.

The rest of the paper is organized as follows: Sect. 2 introduces the related work of sentiment analysis and Chinese micro-blog sentiment analysis. Section 3 introduces the collection and cleaning of data and proves the validity of data. Section 4 introduces the algorithm of the construction of corpus and the training of classifier. Section 5 introduces the vertical and sequential sentiment analysis of micro-blog topics. Section 6 concludes the whole paper.

2 Related Works

Research on the sentiment classification of micro-blog can be divided into two categories: sentiment classification based on sentiment dictionary and sentiment classification based on feature selection and corpus.

The research based on sentiment dictionary focuses on the creation and expansion of sentiment dictionaries. Ku et al. [3] conducted emotional mining of news and micro-blogs and generated the NTUSD sentiment dictionary. Zhu et al. [4] calculated the similarity of words based on HowNet. Sentiment analysis based on feature selection and corpus mostly use machine learning methods to train classifiers. The most representative research is Pang et al. [5] using Naive Bayes, SVM and Maximum Entropy classifier to classify sentiment in English movie reviews. The earliest research on sentiment analysis based on micro-blog short text is distant supervision method proposed by Go [6] in 2009. The method of distant supervision, which applied the method

proposed by Read [7] to construct classified datasets using Emoji, constructed a sentiment dataset containing 1.6 million tweets and achieved a classification accuracy of 83%. Based on this, Pak et al. [8] proposed a method for building a sentiment analysis corpus automatically. Iosifidis et al. [9] built a larger dataset on this basis and applied self-learning and collaborative training methods to expand the dataset.

In recent years, the machine learning method has gradually become the mainstream sentiment classification method. The proposed word2vec [10] for distributed word vector training implements automatic extraction of text features. Especially in the study of micro-blog short text with an average length of only 70, the distributed word vector has the edge over the traditional methods. fastText [11], open-sourced by facebook, can implement unsupervised word vector training and supervised classifier training to train word vector and classifier comparable to deep learning methods in time. As far as we know, there is no precedent for applying fastText to large-scale Chinese micro-blog datasets for training and classification.

The dataflow of micro-blogs is of real-time and timeliness, only by grasping the timeliness of micro-blog information and analyzing the latest topic data can we make even greater use of the value of the data. Culotta et al. [12] monitored the influenza epidemic by analyzing real-time Twitter data. Nahar et al. [13] realized real-time monitoring of cyber bullying based on the Twitter platform. Paul [14] exploited Twitter to construct a spatio-temporal sentiment analysis system to analyze the sentiment trends of voters during the US election. However, most of the researches on analysis of micro-blog focus on the deep learning method to improve the classification performance and apply the typical Stanford Twitter sentiment analysis datasets¹ to train, and there is no vertical and sequential analysis for a particular topic or field in micro-blog. This paper focuses on the sequential study of micro-blog data, seize the timeliness of the data to maximize the value of micro-blog.

3 Data Acquisition and Cleaning

This section will introduce the acquisition and cleaning of Sina Weibo data, concentrate on the design of the general crawler and specialized crawler, the process of data cleaning, and the validity of the micro-blog dataset.

3.1 Data Acquisition

The first way to request micro-blog data is the public interface² provided by Sina Weibo. The interface can only obtain the latest 200 micro-blogs of related topics, in this paper, the data of Sina weibo are collected by web crawler.

We designed two different crawler programs. The general crawler, which applies multi-thread and proxy to achieve the concurrent crawl of 580 thousand micro-blogs per day, is used to collect a large amount of weibo data. The specialized crawler

¹ <https://www.kaggle.com/kazanova/sentiment140>.

² <http://open.weibo.com/wiki/2/search/topics>.

focuses on a certain topic and achieves the micro-blog text within a specific period on the specific topic. From December 1, 2017 to January 31, 2018, the general crawler crawled 35 million micro-blogs in two months, with a total size of 6.34 GB. At the same time, specialized crawler collected micro-blog topic texts from several famous companies from January 2018 to May 2018 for dynamic and sequential sentiment analysis as shown in Table 1.

Table 1. Number of microblogs collected on each topic

| ZTE | MI | MEIZU | HUAWEI | LeEco | Apple |
|--------|--------|--------|--------|--------|--------|
| 38,000 | 34,000 | 59,000 | 58,000 | 73,000 | 47,000 |

3.2 Data Cleaning

This section conducts micro-blog cleaning of datasets, including the cleaning of unique attributes of Weibo, the cleaning of links, mailboxes, special characters and short micro-blogs. The size of the dataset after each step of cleaning is shown in Fig. 1.

First, special symbols like @ and ## are filtered from dataset. Second, statistics show that 0.67 million texts in the micro-blog dataset contain url links or email address, that is, for every 100 micro-blogs, there are two micro-blogs that contain links and email address. We use regular expressions to match and filter url links and email address. Thirdly, special and meaningless characters are filtered in the preprocessing stage. The last step is the filter of short micro-blogs. We stipulate that the micro-blog with length of less than 5 is an invalid micro-blog, in which one Chinese character is counted as one character. After the cleaning of short micro-blogs, 2.28 million invalid short micro-blogs were filtered out, and finally there were 33.48 million valid micro-blogs left in the dataset.

3.3 Validity of Data

After the data cleaning, the jieba³ word segmentation tool was used to segment each micro-blog and generate a dictionary with the size of 649,334. The distribution of word frequency in the dictionary is shown in Fig. 2. We can see from the picture that the number of words that appear once are more than that of twice, and the number of words that appear twice are more than that of third, and so on. This indicates that the distribution of word frequency follows Zipf's Law, which proves the validity of the dataset from the perspective of word frequency distribution.

We also test the validity of the data collected by specialized crawler according to chronological order. Taking "ZTE Topic" as an example, the daily micro-blog statistics from March 21, 2018 to April 27, 2018 are shown in Fig. 3. On April 16th, the number of micro-blogs on ZTE's topic has soared, which corresponds to the incident that ZTE sanctioned by the US Department of Commerce. The validity of the hourly data has also been tested, as shown in Fig. 4. We selected ZTE topic's micro-blogs on March

³ <https://github.com/fxsjy/jieba>.

13, 15 and April 16 for statistics. The number of micro-blogs on March 13, 15 is normal, however, the curves at noon, afternoon and evening with three peaks, which correspond to the three peak periods of people’s landing of Sina weibo. The number of micro-blogs in April 16 was not unusual at noon and afternoon, however, after 9:00 pm, the news that ZTE was sanctioned became a hot spot and the number of micro-blogs increased rapidly. In summary, the data collected by specialized crawler is time-sensitive, which can reflect the public’s attitude towards a topic and can be used for sequential analysis.

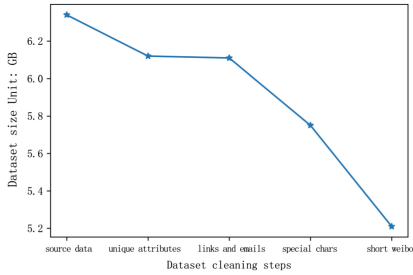


Fig. 1. Cleaning steps of dataset

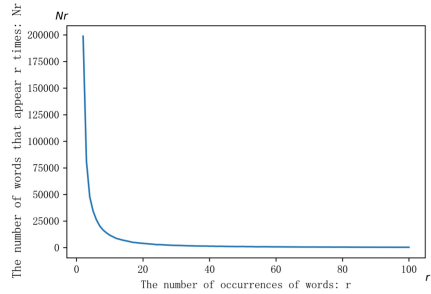


Fig. 2. Word frequency distribution

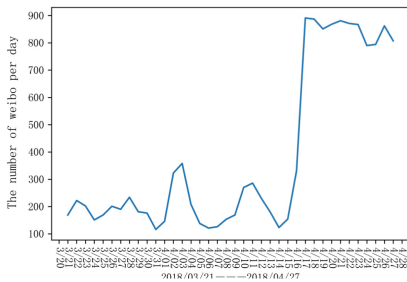


Fig. 3. Number of ZTE weibo from March 21 to April 27

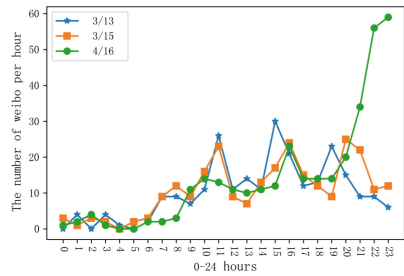


Fig. 4. The 24-hour weibo number of ZTE

4 Corpus Construction and Classifier Training

Section 3 introduces data collection and cleaning and verifies the validity of the data in the end. This section will mainly introduce the construction of corpus and the training of classifiers.

4.1 Corpus Construction

In this paper, we divide people’s sentiments into positive sentiment and negative sentiment. Correspondingly, micro-blogs are divided into positive and negative micro-

blogs, and sentiment-neutral micro-blogs will not be discussed. Chinese characters are extensive and profound, in many cases, it is difficult to tell whether a micro-blog’s sentiment is neutral or not. The third law of Weibo Arguing Laws [15] also points out that people only have positive and negative parties when they quarrel.

The accuracy of manual classification is difficult to control, especially in the condition of big data of 33.48 million micro-blogs. Therefore, we adopt the method of distant supervision learning, using micro-blog Emoji and Sentiment Lexicon to extract positive and negative micro-blogs.

Emoji can be used to express the user’s emotion in micro-blog, to a large extent, it represents the sentiment tendencies of micro-blog text. If a micro-blog contains emojis such as [angry] and [sad], then this micro-blog has a negative tendency. After statistics of dataset, we found that 11.79 million micro-blogs have emojis, accounting for 33.5%. That is, one out of every three micro-blogs have emojis.

















Xie et al. [16] first apply emojis for Chinese weibo sentiment classification, however, they did not consider the ambiguity of certain emojis and mapped all Weibo emojis into positive emojis and negative emojis. Taking the common emoji [smile] as an example, two micro-blogs with [smile] emoji were selected from the Weibo dataset. As shown in Table 2, the sentiments of two micro-blogs are converse. In fact, the emoji [smile] is more often used in young people’s groups to express negative sentiment, so this emoji is ambiguous and cannot be used as a positive flag.

Table 2. Examples of Chinese micro-blog containing ambiguous emoji

| |
|----------------------------------|
| 表情词: [微笑](in English [smile]) |
| 微博消息: |
| 1. 不想和你理论, 你开心就好[微笑] |
| 2. 今天终于见到了我家爱豆, 手动笔芯[微笑][微笑][微笑] |

We manually selected feature emojis with strong sentiment, including 18 typical negative emojis and 37 typical positive emojis. Table 3 lists part of selected emojis. The NTUSD sentiment dictionary is also used as a double filter for Weibo. The detailed filtering process of sentiment micro-blog is shown in Algorithm 1.

Table 3. Typical examples of feature emojis

| <i>Positive emojis</i> | | | | <i>Negative emojis</i> | | | |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Algorithm 1. Emotional Weibo Corpus Construction Algorithm

input: Emotion dictionary NTUSD, Weibo data set *weibos*, Weibo expression dictionary *emoji_dict*

output: Positive weibo collection *pos_set*, Negative weibo collection *neg_set*

1. $pos_set \leftarrow \emptyset$
2. $neg_set \leftarrow \emptyset$ // init set
3. $pos_emotions, neg_emotions \leftarrow load(NTUSD)$
4. $pos_emojis, neg_emojis \leftarrow load(emoji_dict)$
5. **for** *weibo* **in** *weibos* **do**
6. $words \leftarrow set(weibo)$
7. **if** $len(words \& pos_emojis) > 0$ **and**
 $len(words \& neg_emotions) == 0$ **and** $len(words \& neg_emojis) == 0$ **then** // & intersection
8. add *weibo* to *pos_set*
9. **end if**
10. **if** $len(words \& neg_emojis) > 0$ **and**
 $len(words \& pos_emotions) == 0$ **and** $len(words \& pos_emojis) == 0$ **then**
11. add *weibo* to *neg_set*
12. **end if**
13. **end for**
14. **return** *pos_set, neg_set*

The corpus generation algorithm extracts 4.2 million positive micro-blogs and 680,000 negative micro-blogs from the dataset. To prevent the classifier from assigning a bigger priori probability of positive sentiment to the micro-blog in the process of training, we randomly select the same number of micro-blogs as the negative micro-blogs from the positive micro-blogs set, which constitute a corpus of Chinese micro-blog for sentiment analysis.

4.2 Sentiment Classifier Training

In this section, we will apply fastText to train distributed word vectors and sentiment classifiers and introduce some tricks for training sentiment classifier.

Statistics show that the average length of Sina Weibo text is 70, and the short text Weibo with a length of less than 30 accounts for 26%. Short text has the characteristics

of large amount of information and close context. fastText considers context and N-gram features at the input layer and can quickly train and classify short texts. fastText has two efficient word vector training models: CBOW and Skip-gram. Both are three-layer shallow neural network models. The difference is that the CBOW model predicts the current word w_t on the premise of context $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$, Skip-gram model predicts its context $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ on the premise of the current word w_t , and its maximum likelihood function is shown in formula (1). This paper adopts a better Skip-gram model to train word vectors and classifiers.

$$L_{\text{Skip-gram}} = \sum_{w \in C} \log p(\text{Context}(w)|w) \quad (1)$$

Where C is the corpus and Context is the set of context words for word w .

Considering that stopwords will affect the training of word vectors, we trained the word vectors in two kinds of training sets: training set with stopwords and training set without stopwords. The training of the distributed word vector model considers the context of the words, so we speculate that there is no need to eliminate stopwords in the Chinese text word vector training. We also stripped off the emojis from each microblog to prevent fastText from using emojis as a feature in the process of training the classifier.

The experiment results are shown in Fig. 5. When training datasets with 34 million vocabularies and a dictionary size of 360,000, fastText took less than 100 s to implement the training of each word's 100-dimension word vector. The classifier achieved more than 90% accuracy on the test set, and the classifier based on the training set with stopwords was 0.4% higher than the classifier trained without stopwords, which proved that the stopwords play an important role in the training of distributed word vectors. We also conducted 200-dimension and 300-dimension word vector training, as the training time doubled, the accuracy of classifier only increased by 0.1%. The classifier with the highest accuracy is the 300-dimension classifier trained with stopwords, which achieves an accuracy of 92.2% compared to the 83% achieved by Alec Go's maximum entropy method.

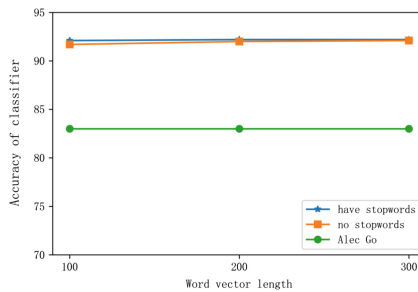


Fig. 5. Accuracy of word vector classifier

5 Experiments

In this section, we apply sentiment classifier trained in Sect. 4 to perform dynamic and sequential sentiment analysis of micro-blog topic data collected in Sect. 3.

Considering that ZTE has a hot topic in the first four months of 2018, this section focuses on a detailed analysis of ZTE’s topic. In Sect. 3, we collected 38,000 micro-blogs of ZTE, with an average of 310 micro-blogs per day and 13 micro-blogs per hour. Micro-blog with a probability greater than 0.7 is marked as corresponding positive or negative micro-blog, which can be used to further filter the noise micro-blogs. Finally, the sentiment classification results of each micro-blog are obtained in chronological order. As shown in Fig. 6, a sequential sentiment analysis of ZTE Topics from March 21 to April 27, 2018 was conducted in units of days. From the diagram, we can intuitively see that the topic sentiment from March 21 to April 15 are all positive. Until the incident in which ZTE was sanctioned by the US Department of Commerce on April 16, negative micro-blogs increased sharply and exceeded positive micro-blogs, reflecting the huge crisis faced by ZTE. We also conducted the more vertical analysis of the 24 h on April 16th. It can be seen from Fig. 7 that the number of positive micro-blogs exceeds the negative micro-blogs in most of the time. After 21 o’clock, negative micro-blogs rose suddenly and continued to increase, which corresponds exactly to the “ZTE Denial Order” at 9:00 a.m. in US.

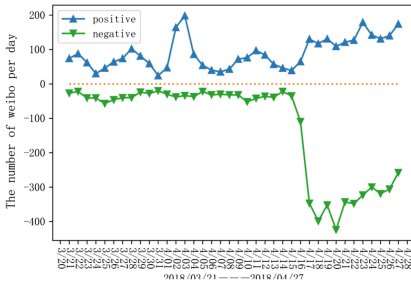


Fig. 6. Number of positive weibo and negative weibo from 3/21 to 4/16 about ZTE

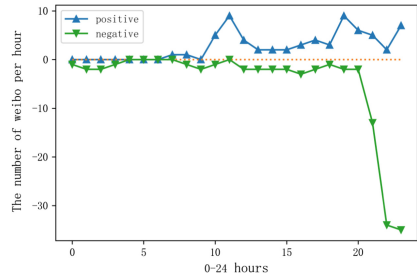


Fig. 7. Number of positive weibo and negative weibo in 4/16 about ZTE

The analysis above proves the validity of the vertical and sequential analysis of Sina Weibo topics. Different from the degree of heat provided by Baidu index and Weibo micro index, the analysis method realized in this paper not only considers the popularity of topic, but also realizes the dynamic sequential analysis of positive emotions and negative emotions, which can quickly and intuitively reflect the sentiment changes of the topic.

6 Conclusion

The vertical and sequential sentiment analysis of micro-blog has broad application prospects. This paper collects data from Sina Weibo and proposes a distant supervision learning method based on Emojis and Sentiment Lexicon. A large-scale Chinese micro-blog sentiment analysis corpus is constructed, and fastText is used to train word vectors and sentiment classifiers. Experiments show that the classifier surpasses traditional classifier in performance and can accurately reflect the sentiment trend of a topic in practical applications. Limited by the length of the article, the results of other Weibo topics, manually selected emojis and ZTE topics' Weibo data will be shared on Github⁴. Next, we will focus on the classification of neutral emotion micro-blog and apply the research results to the public opinion analysis and early warning of public sentiment.

References

1. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Meeting of the Association for Computational Linguistics: Short Papers, 08–14 July 2012, Jeju, Island, Korea, pp. 90–94. Association for Computational Linguistics (2012)
2. Joulin, A., Grave, E., Bojanowski, P., et al.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers, April 2017, pp. 427–431. Association for Computational Linguistics (2017)
3. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: AAAI-CAAW (2006)
4. Zhu, Y.L., Min, J., Zhou, Y., et al.: Semantic orientation computing based on HowNet. *J. Chin. Inf. Process.* **20**(1), 14–20 (2006)
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of EMNLP, pp. 79–86. Association for Computational Linguistics (2002)
6. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *Cs224n Project Report* (2009)
7. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. *Knowl. Inf. Syst.* 1–47 (2018)
8. Park, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010, Valletta, Malta. *DBLP* (2010)
9. Iosifidis, V., Ntoutsi, E.: Large scale sentiment learning with limited labels. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1823–1832. ACM (2017)
10. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. *Comput. Sci.* (2013)

⁴ https://github.com/wansho/Weibo_Sentiment_Analyze.

11. Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching word vectors with subword information (2016)
12. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics, Washington, DC, Columbia, 25–28 July 2010, pp. 115–122. ACM (2010)
13. Nahar, V., Al-Maskari, S., Li, X., Pang, C.: Semi-supervised learning for cyberbullying detection in social networks. In: Wang, H., Sharaf, M.A. (eds.) ADC 2014. LNCS, vol. 8506, pp. 160–171. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08608-8_14
14. Paul, D., Li, F., Teja, M.K., et al.: Spatio temporal sentiment analysis of US Election what Twitter says!. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1585–1594. ACM (2017)
15. Du, J.F.: Weibo arguing laws: a sketch of social psychology on the Internet. *News Writ.* (06), 65 (2017)
16. Xie, L., Zhou, M., Sun, M.-S.: Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction. *J. Chin. Inf. Process.* **26**(1), 73–83 (2012)