Sentiment Analysis of Weibo Comments based on LDA Model

Chen Fang

Wenzhou No.2 Foreign Language School, Wenzhou, China

2219088342@qq.com

Abstract. The essence of LDA (Latent Dirichlet Allocation) model is a generative Bayesian probability model that contains three layers of words, topics and corpus (sometimes called document set). Under the LDA algorithm theory, each document represents a probability distribution formed by some topics, and each topic represents a probability distribution formed by many words. Therefore, the model fitting results will present the core keywords and specific probabilities of each topic, and researchers can interpret the meaning of the document according to the model results. In this paper, we hope to use the LDA model as the basis for the emotional analysis of microblog comments.

Keywords: LDA Model; Weibo; Sentiment Analysis.

1. Introduction

Emotional analysis of microblog comments refers to the analysis of the opinions and attitudes expressed in microblog comments. The main purpose is to judge the subjective and objective, positive and negative polarity and emotional tendency intensity reflected by the publishers. Chinese microblog sentiment analysis involves natural language processing, information retrieval, machine learning and other fields, and can be applied to hot event discovery, psychological analysis, network public opinion management and other tasks [1].

With regard to the research results of Chinese text emotion analysis, there have been studies from the perspectives of syntax analysis, knowledge engineering, machine learning, etc[2]. to explore the text orientation in the target field. This paper takes the comments on hot topics in microblog as the research object, analyzes the characteristics of comment texts in terms of word use and emotional expression with the help of LDA model [3], and obtains the characteristics of the spread, development and change of netizens' emotions in this event, which can provide reference for public opinion disposal in related fields.

As one of the most influential social interactive new media nowadays, microblog has a lot of data related to user privacy [4]. It authorizes developers to obtain some data on the platform in a specific way for their research and analysis. At present, there are two main ways to crawl microblog comments: one is to capture them through the API interface provided by microblog officials, and the other is to obtain them through parsing microblog pages. The second way can be divided into three basic steps: a. Select the appropriate browser, Google here, and use the mobile terminal's website for microblog related access; b. Decompose the obtained request list and find the source url of structured json; c. Analyze the rules of url and write corresponding python code [5]. This paper adopts the second method, using LDA model for analysis.

2. Research Design

First of all, this paper compiles Python multi-threaded crawler code based on Selenium, crawls the microblog comments that can be obtained in nearly 180 days, and collects the release time, publisher, comment content and other information of all comments to form the original text dataset. Secondly, use Python to preprocess the text data set, such as removing useless characters such as emoji, advertising logo, invalid link and address, and converting English uppercase letters to lowercase, and Chinese traditional characters to simplified characters. Redundancy is deleted through text de duplication and mechanical compression, and then short text lacking practical effect is eliminated. Finally, we borrowed the Jieba thesaurus [6], combined with the existing corpus to form a user

segmentation dictionary, segmented the microblog comment dataset, and called the stop dictionary of Harbin Institute of Technology to remove stop words to obtain the basic corpus for subsequent LDA modeling [7].

In topic extraction and analysis, this paper converts the preprocessed corpus into a bag sparse matrix, calculates the corresponding TF-IDF value, observes the contribution of each word item to the document topic, and then filters high-frequency and low discrimination words to obtain the document feature vector [8]. The document feature vector is used as the input of LDA model to establish the topic model, and the output results of the model are manually encoded to obtain various information topics [9].

In the experiment, Python was used to preprocess the collected data, and 1984662 original word segmentation corpora were obtained for LDA topic modeling. By building the word bag model, a total of 87742-word dictionaries and corresponding word bag text vectors are established. On this basis, TF-IDF model in the gensim library is called to calculate TF-IDF values of the document sparse matrix as the text feature vector [10].

3. Empirical Conclusion

Due to the huge amount of micro blog comment data, we choose a typical example as the conclusion for analysis. The experimental data comes from Sina Weibo topic - four PLA soldiers died in the border conflict between China and India. This topic has been on the hot search ranking since February 19, 2021, and has received tens of thousands of comments as of April 1. Use Octopus software to crawl the comment data of topic news in the next two weeks. After deleting duplicate and contentless comments, 4177 valid data are obtained, and the link address, "@ character" symbol and "# topic" text sequence in the data are filtered. Download PKuSeg's online text model of Peking University to segment and tag the comments. With the help of Harbin Institute of Technology's stop words list, stop words and comments with only emoticons are removed. Finally, with the user's name as the index, the segmentation results are stored in the database to complete the preprocessing of the original data.

High frequency emotional words refer to the emotional words frequently used in comments, which, to a certain extent, represent the main emotional tendency of netizens in the event. The word frequency and polarity of all comments were counted and analyzed, and the top 6 positive and negative words were obtained. It can be seen that the frequency of the words "hero" and "salute" is far ahead, which expresses the great respect of many netizens for the heroes who sacrificed, thanks them for their protection and brings peace, and calls on everyone to remember the history, reflect on the current situation and cherish the good life at present. The word "peace" reflects the emotion and reflection of netizens in the context of the event.

In the high-frequency words of negative emotions, netizens used such words as "blood debt and blood compensation" and "aggression" to show their attitude towards foreign troops. The word "forget" is used against its meaning, emphasizing that we should not forget why the soldiers died. These words, on the one hand, show netizens' feelings of sorrow and regret for the lost heroes, and on the other hand, express that he was dissatisfied and indignant with the actions of foreign troops, showing obvious emotional characteristics. Simply relying on dictionaries to calculate the emotional value may lead to the situation that the comments themselves contain commendatory and derogatory meanings, but the ultimate value is 0 due to insufficient or no emotional words. Therefore, it is necessary to further explore the content composition of network text and summarize the characteristics of emotional expression.

Analyzing the comments with an emotional value of 0, we found that there are four main situations. In these four cases, the number of English comments is small and repetitive. For example, "salute" and "inspect" are online expressions of respect, and "CNCN" expresses national pride.

The rhetorical questions mainly focus on negative emotions, mainly involving the details of the report that netizens pay attention to, the follow-up progress of the event and other issues, and convey

more intense emotions [11]; Although there are also words such as "martyr" and "husband" in the poetry to assist in emotional judgment, there are ancient and modern synonyms and polysemy of a word, and the sentence reading is also different from today's participle.

From the perspective of emotional expression, in addition to the courage of sacrificing soldiers to protect mountains and rivers, there are also reflections on the social status quo; Relatively speaking, emotional sentences rarely involve emotional words, and the expression is not direct enough, which requires some background knowledge for reasoning. It is difficult to achieve the effect of text understanding only by syntax and word matching, which requires manual evaluation.

According to the method mentioned above, calculate the emotional value of the above 4177 comments. If it is greater than 0, it will be recorded as a positive emotion, otherwise it will be recorded as a negative emotion. Finally, 3053 positive emotions were obtained, accounting for 73.1%; 1124 negative emotions were obtained, accounting for 26.9%. On the whole, rational voice accounts for the majority, but negative comments still account for a certain proportion.

The network public opinion event of China India border conflict caused heated discussion after it was disclosed by Phoenix. com, and then the popularity gradually declined. In order to discuss the emotional change trend of netizens in two weeks, after determining the polarity of each comment, the proportion of positive and negative comments in two weeks is counted, and the time change trend of the emotional proportion is analyzed with the timestamp as the index. It can be seen that on the first day of the news release, netizens showed some negative emotions. On the afternoon of February 19, 2021, the Ministry of National Defense issued a response to the issue of the intention of the officers and soldiers to defend the country and defend the border. On the evening of February 19, the contents of martyr Chen Xiangrong's diary were made public, the attitude of netizens was eased, and the proportion of positive emotions began to increase. On February 20, netizens who denigrated the frontier defense heroes were detained criminally. Meanwhile, some big We Media networks spread the positive energy of the army, playing a positive role in guiding public opinion. In the following days, with the release of the video on the scene of the conflict, the progress of the negotiations between the two armies, and the related reports of the martyrs' families, netizens' emotions fluctuated to varying degrees. On March 1, "the first lesson of school" themed with honoring the frontier heroes was held in many places, leading public opinion to healthy development. Since then, with the passage of time, the heat of the event has gradually faded, and the proportion of positive and negative emotions has gradually stabilized. According to the periodic evolution theory of online public opinion events, the emotional evolution can be divided into four stages: beginning, diffusion, climax and regression, which have obvious stage characteristics.

The characteristics of unconventional online public opinion are different from those of conventional public opinion. It almost reached a climax at the beginning, and then gradually subsided. From this point of view, the evolution of public opinion represented by the Sino Indian border conflict is more in line with the characteristics of unconventional public opinion. This requires the government to more closely track the trend of public opinion, strengthen media control and give timely response. For example, after the PLA announced the casualties, some netizens questioned the reasons for the official reports every six months, and the Ministry of National Defense responded quickly to allay public doubts; Some man-made rumors with ulterior motives caused trouble, and the public security department took timely action to avoid the spread of malicious comments. Within two weeks of the release of the news, various kinds of reports alternated, and the official media timely guided, cooperated with opinion leaders with high credibility, actively publicized positive events online and offline, jointly promoted the theme and positive energy, and gradually guided the trend of public sentiment.

4. Conclusion

The information flow in the network era is increasingly smooth, and microblogging platform, as a public space, has become a gathering point of various emotions under emergencies. The emotional

analysis of microblog comments is helpful to grasp the law of public opinion development and provide reference for government decision-making. This paper takes microblog hot topic comments as the data base, integrates multiple emotion classification dictionaries, extracts high-frequency emotional words, combines LDA model, analyzes the emotional polarity of the comment text, obtains the proportion of positive and negative emotional comments, and discusses the temporal change of netizens' emotions within two weeks. It is found that the comments with positive emotional tendencies occupy the majority. The close tracking and timely response of the media also play a positive role in guiding public opinion.

Compared with the standard text, the review text is shorter, and the relationship and dependence between emotional words are smaller. Therefore, the research scheme based on emotional dictionary is more suitable. This experiment verifies the feasibility of the method and can be used as a reference and improvement idea. At the same time, it should also be noted that the emotional analysis itself is highly subjective, and the network text expression forms are also constantly changing and updating. When the topic involves multiple fields or the text structure is complex, it is necessary to combine the dictionary and supervised learning methods to solve the problem by integrating the advantages of the two ideas. In the next step, we should try to introduce word similarity calculation and relevant machine learning algorithms to better grasp the text content and improve the accuracy of calculation.

References

[1] Zhang C, Sun J, Ding Y. (2011). Topic Mining for Microblog Based on MB-LDA Model. Journal of Computer Research and Development.48(10):1795-1802.

[2] Li W B, Le S, Zhang D K. (2008) Text Classification Based on Labeled-LDA Model. Chinese Journal of Computers.31(4):620-627.

[3] Shi J, Wan-Long L I. (2010). Topic Words Extraction Method Based on LDA Model. Computer Engineering.

[4] Wang L, Wei B, Jie Y. (2011). Topic Discovery based on LDA_col Model and Topic Significance Re-ranking. Journal of Computers.6(8):1639-1647.

[5] Liang J, Liu P, Tan J, et al. (2014). Sentiment classification based on AS-LDA model. Procedia Computer Scienc.31:511-516.

[6] Gao Y, Chen J, Zhu J. (2016). Streaming Gibbs Sampling for LDA Model.

[7] Tang Z, Yuan W. (2021). Marketing Improvement of Chinese Original Picture Books from Dissatisfaction Evaluation - Text Mining Based on LDA Model. SHS Web of Conferences. EDP Sciences.

[8] Wang Z Z, Ming H E, Yong-Ping D U. (2013). Text Similarity Computing Based on Topic Model LDA. Computer Science.

[9] Yu C. (2010). Mining Hot Topics of User Comment Based on LDA Model: Principle & Approach. Information Studies: Theory & Application.

[10] Zhang H, Li S, Feng J, et al. (2021). Public Opinion Analysis of Weibo Comments Based on Crawler and SVM.2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE.

[11] Luo C, Dan T, Li Y, et al. (2019). SENTIMENTAL ANALYSIS OF WEIBO COMMENTS BASED ON FAKE COMMENTS RECOGNITION AND ITS APPLICATION. Computer Applications and Software.

[12] Fang W. (2019). Sentiment Analysis of Weibo Comments Based on Deep Neural Network.2019 International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE.