Research on Sentiment Analysis Model of Douban Book Short Reviews based on Machine Learning

Zhen Zhang

Fuzhou University, Fuzhou, China

Abstract. The era of big data has produced a large amount of text data. Due to the different data sources and the complexity of the data generation process, a large amount of data has a heterogeneous structure. Combining several latest machine learning methods with hierarchical technology, the algorithm for text data with heterostructure can improve the accuracy of text classification. Empirical data analysis shows that our algorithm has a significant effect on improving classification accuracy. Based on the grounded theory, this paper analyzes 2000 "short comments" of sample books on the "Douban Reading" website, and explores the motivation of short comments and the corresponding emotional analysis. The data results show that users' emotional tendencies in Douban applications are diversified and consistent, and neutral comments play an important role. The user's emotional tendency represents the user's evaluation attitude towards books. The user's emotional tendency in the evaluation can be grasped in a timely manner. Readers can be used to understand the user's recognition of books in a timely manner and adjust their service strategies in a timely manner.

Keywords: Machine Learning; Douban; Short Book Reviews; Sentiment Analysis Model.

1. Introduction

With the development of the Internet, people are used to expressing their feelings by publishing various comments on the Internet, and most of them will collect articles or books that they like better or other websites that they are interested in [1]. Generally speaking, when users buy products online, they will browse the comments of other users to help them make decisions [2]. At the same time, a large number of comments involving consumers have also been generated, providing a good sample for the study of consumer demand characteristics [3]. Taking the short review of Douban books as an example, we can get a lot of valuable review information, which can comprehensively reflect the characteristics of reading needs. Text mining and emotion analysis are carried out for mobile library's social services to discover users' evaluation and emotional tendency to mobile library resources, so as to better realize mobile library resource promotion and accurate recommendation service, which has become a new research hotspot [4]. This research starts with the overall attitude of netizens' comments on Douban, then analyzes the high-frequency words in netizens' short comments, then analyzes the most popular reviews, and finally clusters all the collected short comments. The object of analysis is from attitude, to words, to paragraphs, and finally to the data analysis of all short comments. The whole process is a gradual and in-depth analysis of netizens' impressions.

2. Theories Related to Machine Learning and Sentiment Analysis

2.1 Development of Sentiment Analysis

We need to analyze the target document in more detail. Sentence level emotion analysis task is to determine whether a sentence expresses positive, negative or neutral emotions [5]. This level of emotion analysis is closely related to the subjective and objective classification tasks. The objective and subjective classification task is to distinguish whether the statement is to state factual information (objective sentence) or express subjective information (subjective sentence). Attribute level emotional analysis is a more detailed emotional analysis, which directly focuses on goals and opinions rather than linguistic units such as text and sentence [6]. For researchers, focusing on the extraction and analysis of evaluation objects can better understand the problem of emotion analysis. According to the current research, we can draw a frame diagram of emotion analysis, as shown in Figure 1 below:

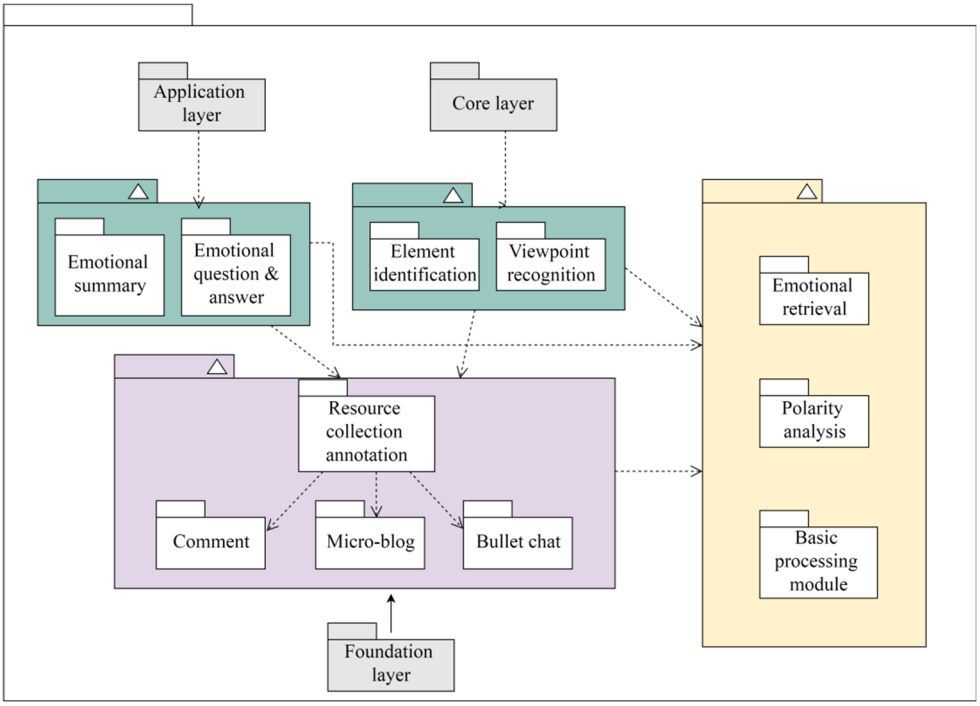

Figure 1. Sentiment analysis model framework

Emotion analysis is to use machine learning or deep learning algorithm, combined with common methods of natural language processing to classify a paragraph of text, and its essence is text classification. It is a sub-direction in the field of natural language processing. At present, affective analysis can be divided into three levels: discourse level, sentence level and attribute level [7]. The main purpose of comment sentiment analysis is to predict the emotional polarity of comments, such as positive comments and negative comments [8]. Emotion analysis has an important application in e-commerce, social media analysis and marketing, making it a research hotspot [9]. An emotion analysis model with high accuracy can not only save a lot of manpower and material resources, but also create a lot of economic and social values [10].

With the rapid development of all kinds of new media, new words on the Internet emerge in endlessly, and the meaning of existing emotional words is also changing constantly, which may even produce completely different polarities in different contexts [11]. Therefore, how to find new words more efficiently and expand existing dictionaries, while accurately judging the polarity of related words in different contexts and optimizing the analysis model, will be very important for decision-making based on the results of emotion analysis. Combined with book review theme modeling and emotion analysis, the same attention attribute and its review content of different books can be displayed side by side, and further set up targeted functional display areas to achieve book promotion and content innovation.

2.2 Classification of Sentiment Analysis Models based on Machine Learning

Machine learning refers to knowing the subsequent behavior according to the acquired empirical data, in which the subject of acquiring empirical data and the object of knowing the subsequent behavior are both machines. In order to know the subsequent behavior, after the machine obtains the empirical data, it must conduct data analysis to obtain the model information contained in the data and make behavior decisions based on the information. The process of analyzing empirical data can be seen as learning. Compared with the method of emotion dictionary, machine learning method is simpler, more accurate for emotion analysis, better scalability and repeatability, and can achieve higher classification accuracy. However, the classification accuracy of machine learning depends on high-quality labeled training sets. Large scale high-quality training data requires high labor costs, and human subjective data labeling results will also affect the classification effect.

The traditional method based on dictionary and rules to classify the emotional polarity of book evaluation is to divide texts into three categories: positive emotion, neutral emotion and derogatory emotion. Firstly, the emotional information of words is modeled. The cross-domain affective word vector generation model adopts the method of machine learning, so that the word vector model can predict the emotional tendency of words while using context to predict the central words. Predicting the emotion of words can be regarded as an emotion classification problem. By adding the loss function of emotion prediction into the model, and training and adjusting the word vector to minimize the loss function, the emotion information can be integrated into the word vector. In addition, word vectors are trained on large corpus data in an unsupervised way. How to effectively use data from different fields to train word vectors remains to be solved.

3. Empirical Research

When users realize that the book is not as good as they imagined or expected after reading, or they disagree with the arguments in the book to alleviate regret or disappointment, they will express their unhappy reading experience by publishing book reviews on Douban. The motivation of users to generate short comments includes six main categories: discussion, sharing, entertainment, external pressure, emotional venting and data acquisition.

According to the usage habits and grammatical structure of Chinese, this paper selects a five-part graded lexicon of adjectives, verbs, adverbs, nouns and idioms, with 200 words and idioms, 500 adjectives, 80 adverbs, 200 nouns, 365 idioms and 154 verbs. In order to reflect the emotional tendency of this review text more accurately and carefully, this paper divides nouns, adjectives, idioms and verbs into four emotional levels, which are -5, 0, 5 and 10 respectively. Thus, the emotional score calculated in this paper is a continuous score, so that the degree of emotional tendency can be well reflected in the text, and our understanding of this kind of comment text can be more specific and accurate (see Table 1).

Table 1. Sentiment dictionary example

Adv	score	Tone	score
Super	1.3	No	-1
Very	1.6	Whatever	1
Incomparable	1.2	Doesn't it	1
Especially	0.79		
Slightly	0.84		
A bit	0.65		

The data obtained after word segmentation cannot be directly used as the processing text for subsequent work, so it is necessary to process and filter the text data after word segmentation. With the increase of text length, the number of model parameters will grow too fast. If the text is too long, it may lead to too many parameters, which will sharply degrade the classification performance of the model. However, at this stage, the data scale is growing, and the text to be predicted is also growing.

In order to analyze the emotion of text data, a key basic operation is to select an emotion lexicon suitable for this study. When calculating the emotional score of a sentence, the emoticon of a single positive (negative) emotion is 1 (- 1) point, and the emoticon of the same consecutive positive (negative) emotion is 2 (- 2) points. According to the emotional vocabulary and the corresponding emotional polarity intensity in each comment, the formula for calculating the emotional score of this comment is:

$$s = \sum_{i=1}^{k} s_i (k-1)^{N_i} + \sum_{j=0}^{j} e_j \quad (1)$$

Among them, s is the final emotional score of the comment, which includes k emotional words and j emoticon groups. s_i is the polarity intensity of the emotional word in the comment, and there is a total of N_i negative words modifying the emotional word. The first k nearest word like word vectors control the moving direction of the target word vector, while w_i controls the moving distance of the target word vector, that is, how far the target word vector will move towards its approximate word vector.

Find the matching emotion words in the constructed domain emotion dictionary, and take each emotion word matched as the benchmark, find the degree adverbs and negative words in turn forward, and calculate the corresponding score. Sum the scores of each emotional word in the clause, and finally accumulate the scores of all clauses of the user comment to obtain the final score of the user comment (Table 2).

Table 2. The experimental results of selecting the best value of the number of approximate words

Corpus	Number of synonyms	P	R	F
Douban-10c data set	5	0.62	0.73	0.65
	15	0.67	0.62	0.64
	20	0.79	0.64	0.61
	25	0.77	0.73	0.77
	30	0.73	0.83	0.84
Douban-102d dataset	5	0.63	0.83	0.79
	15	0.6	0.66	0.78
	20	0.79	0.7	0.64
	25	0.67	0.77	0.8
	30	0.67	0.71	0.73
Chn htl_all dataset	5	0.66	0.63	0.81
	15	0.6	0.61	0.67
	20	0.73	0.63	0.83
	25	0.74	0.7	0.6
	30	0.6	0.75	0.7

Suppose S is a sentence with practical meaning, including a series of words $W_1, W_2, W_3, \cdots, W_n$ in a specific order. Where n represents the length of the sentence, that is, the number of words in the sentence. W_i^j represents the text segment from the i th word to the j th word in statement S. Then the possibility of sentence S can be expressed as:

$$P(S) = p(W_1^n) = p(W_1, W_2, W_3, \cdots, W_n) \tag{2}$$

To sum up, the sentiment analysis model for machine learning proposed in this chapter outperforms other comparative models in the NLP (Natural language processing) aspect-level sentiment analysis task. The emotional tendencies of users in the Douban application are diverse and consistent, and neutral comments play an important role. This shows that the method of sentiment analysis of readers' comments is helpful to identify niche books that are not recommended by the platform, and can undoubtedly increase users' enthusiasm for books. The use of online book platforms can not only enrich website content, but also increase user stickiness.

4. Conclusion

Through the generation, sharing and dissemination of mobile internet apps, users can freely express their personal views and comments, achieve interaction and communication with other users, and integrate their personal opinions and opinions into the system as a whole. In today's network environment, new words are constantly emerging, and the text information in the network is even more vast, which makes it difficult for traditional methods to fully extract and train text features. When processing large quantities of data, the analysis performance will significantly decline. Therefore, a more efficient and accurate method of emotion analysis emerged - emotion analysis model based on machine learning. Using sentiment analysis technology to analyze comment text automatically can not only solve the problem of soaring cost, high error rate and low efficiency caused by manual analysis, but also create greater social value. Therefore, the research on emotion analysis technology has great practical significance.

References

[1] Li Mengnan, Wang Mingyan. A review of sentiment analysis methods and applications based on machine learning [J]. Software Engineering, 2021, 24(9):4.

[2] Xue Tao. Research on a Python-based Machine Learning Sentiment Analysis Method [J]. Journal of Jiamusi University: Natural Science Edition, 2020, 38(3):4.

[3] Li Ding. Research on the Improvement of Machine Learning Sentiment Analysis Method [J]. Journal of Xi'an Institute of Aeronautics and Astronautics, 2020, 38(1):6.

[4] Li Ding. On the Basic Theory of Machine Learning Sentiment Analysis [J]. Farm Staff, 2020, No.661 (14): 280-280.

[5] Li Ding. Talking about the Machine Learning Sentiment Analysis Method [J]. Farm Staff, 2020, No.648(05):170+205.

[6] Song Zukang, Yan Ruixia, Gu Liqiong. Text topic generalization and sentiment analysis based on machine learning and sentiment dictionary [J]. Software Guide, 2019, 018(004):4-8.

[7] Sun Jianwang, Lv Xueqiang, Zhang Leihan. Research on Chinese Microblog Sentiment Analysis Based on Dictionary and Machine Learning [J]. Computer Applications and Software, 2014, 31(7):5.

[8] Han Bixiang. Sentiment Analysis of "Number Porting" Weibo Comments Based on Machine Learning [J]. Word world, 2021,23(1):36.

[9] Kinyua J D, Mutigwe C, Cushing D J, et al. An analysis of the impact of President Trump's tweets on the DJIA and S&P 500 using machine learning and sentiment analysis[J]. Journal of Behavioral and Experimental Finance, 2021, 29(2):100447.

[10] Kinyua J D, Mutigwe C, Cushing D J, et al. An analysis of the impact of President Trump's tweets on the DJIA and S&P 500 using machine learning and sentiment analysis[J]. Journal of Behavioral and Experimental Finance, 2021, 29(2):100447.

[11] Kalaivani M S, Jayalakshmi S. Sentiment analysis on micro-blog data using machine learning techniques - A Review [J]. IOP Conference Series Materials Science and Engineering, 2021, 1049 (1): 012012.