Text Analysis of Commodity Evaluation Data Mining Based on LDA Theme Model

Qian Zhang(✉)

School of Management, Wuhan Textile University, Wuhan 430073, China
1048564342@qq.com

Abstract. With the development of e-commerce industry, product reviews have become an important bridge between consumers and businesses. Consumers intuitively express their opinions and comments on products or services through product reviews, making online reviews an important basis for online businesses to improve service quality. In order to improve the service quality of online merchants, this paper uses Python's crawler technology to crawl 1000 favorable comments, 1000 moderate comments and 1000 bad comments of JD.COM Midea Heater, and mines and analyzes online comments, and uses LDA (Latent Dirichlet Allocation) model to process online comment data, so as to find out consumers' concerns about goods, thus providing merchants with ideas for improving services from comments.

Keywords: Commodity evaluation · LDA model · Data mining · Consumer preference

1 Introduction

With the gradual improvement of internet infrastructure, the network coverage rate has increased, the scale of network users has expanded, and the network situation has improved. According to the 47th Statistical Report on China's Internet Development released by China Internet Network Information Center (CNNIC) in Beijing, by December 2020, the number of Internet users in China had reached 989 million, an increase of 85.4 million compared with March 2020, and the Internet penetration rate reached 70.4% [1]. According to statistics, China's online retail sales reached 11.76 trillion yuan in 2020, an increase of 10.9% compared with 2019. The continuous development of mobile Internet has become the continuous growth force of online consumption in online shopping, which makes it possible for online information transmission and online supply market of consumer demand groups. Massive online comment data is exploding. Under the shopping environment of e-commerce platform with overloaded network information, useful online comment information has gradually become an inevitable reference for consumers to make online shopping decisions. As the most accessible source of commodity information, online reviews are evaluation information created by users based on their personal consumption experience. As a new element in the marketing information communication combination, online reviews can enable consumers to browse, share

G. Vilas et al. (Eds.): ICEMED 2023, AEBMR 254, pp. 316–320, 2023.
https://doi.org/10.2991/978-94-6463-224-8_39

and spread freely on merchant platforms, online communities and third-party websites for free, and help users to judge and identify their favorite products efficiently, thus becoming the focus of online shopping consumption.

The research status of commodity evaluation can be divided into three aspects: first, explore the consumer's emotion from commodity evaluation [2]. Zhang and others analyze consumers' emotional tendency through data mining to help enterprises better adjust their service quality. Secondly, through deep learning to analyze commodity evaluation [3], Kang and others put forward an emotional analysis model of commodity evaluation based on deep learning [4]. Thirdly, other influencing factors in commodity evaluation, such as pictures and reply information, are studied.

Most of the literature studies commodity reviews from the empirical direction. This paper analyzes the online review data and studies how to improve the service quality of merchants. Select the comment data of heaters in "JD.COM Shopping Mall" to carry out text mining, and carry out theme mining for online comment content of favorable comments, moderate comments and negative comments, trying to find different characteristics of merchant products from the perspective of consumers.

2 Relevant Theoretical Knowledge and Technical Interpretation

2.1 Text Mining

Text mining refers to the application of natural language processing and analysis methods to convert the text into data and then analyze it, with the aim of extracting high-quality knowledge information with potential value from the text.

The main process of comment text mining is as follows: First, get text data. Obtain original text data from text database, Web page or other channels. Second, the text is preprocessed. It mainly includes word segmentation, part-of-speech tagging, stopword removal, feature extraction and other steps. Third, carry out text mining. After the comment text is converted into structured data, the structured data is further analyzed. Fourthly, model evaluation and result display, and visualize the mined information. Text mining is often used in text classification and clustering, machine translation, information filtering and automatic speech recognition, and its application is very extensive. Commodity reviews exist in the form of text, so in the face of such a huge amount of review text in the network environment, it is necessary to use text mining methods and technologies to analyze it. In this way, an effective method is obtained to preprocess and classify review texts and extract product features.

2.2 LDA Theme Model

The LDA theme model is a generation model. That is, it is considered that every word in an article begins with the article, and a specific topic with a specific "document topic" probability distribution and a word probability distribution are selected and repeated continuously until a document is generated [5]. After LDA topic model training, we will get two most important results-the probability distribution of topics under documents and the probability distribution of words under topics. Parameter estimation is

the inverse process of this training process. There are many methods of parameter estimation, mainly variational inference EM algorithm and Gibbs sampling method. Gibbs sampling algorithm is adopted in this paper.

3 Research Process

Because the quantity of goods purchased in JD.COM leads to more comments, and the threshold of evaluation is low, users can express their own comments and opinions after purchase. The data needed in this paper is the commodity evaluation of the famous brand Midea heater in JD.COM Mall, including the evaluation content, product size, score and delivery time. In this paper, the tool for collecting data is to use the crawler mode in python language to crawl the product web page. A total of 1000 reviews, 1000 middle reviews and 1000 bad reviews were obtained. However, the comment data has the characteristics of sparsity, irregular expression, including multiple emoticons and redundancy of big data. In this paper, the crawling text processing mainly uses natural language processing and data mining to clean, filter and classify the crawling comment data, so as to select better quality data and provide a better data environment for subsequent research. Because the research of text focuses on the theme analysis of commodity evaluation, this paper only selects text information for analysis. The topic model does not directly analyze text documents, but analyzes the document word matrix based on these documents, which sets the frequency of each word appearing in the document. First of all, we use python to finish the text preprocessing, and use modules such as "re,nltk,spacy" which are commonly used to deal with natural languages. The specific steps are as follows: (1) Text segmentation, for Chinese, we use the now mature jieba word segmentation tool for word segmentation; (2) stem extraction, which is one of the steps of part-of-speech standardization, extracts the stem or root of a word in a reduced way; (3) Part-of-speech standard, which labels all segmented words as nouns, adjectives, verbs, prepositions, etc. according to their parts of speech; (4) Feature extraction, because the frequency of different words and phrases in the document is quite different, some words only appear a few times or even less in the process of analysis, or some words appear more times, but they have nothing to do with the meaning of the article, so it is necessary to filter these rare words and meaningless words. By setting the minimum frequency of words in the document as 1 as the threshold, those rare words are filtered, and those common and meaningless words are filtered through the commonly used stoplist. At this point, the preprocessing of the text has been completed.

The wordcloud module is called by Python language to generate three word cloud images. The word cloud map of good, medium and bad reviews is shown in Fig. 1.

Perform LDA keyword extraction. LDA is an unsupervised learning model, which can discover hidden topics in documents through document semantics. Its core idea is to express documents abstractly as the probability distribution of topics, and topics are constructed as the probability distribution of vocabulary, thus linking documents and vocabulary. LDA keyword extraction is shown in Table 1. Theme model diagram is shown in Fig. 2.

Fig. 1. Word cloud map of good, medium and bad reviews of goods

Table 1. Collation of LDA keyword extraction results

Key word	Theme
heating, heating, temperature	product performance
design, blades	product appearance
customer service, express delivery, after-sales, price reduction	product service

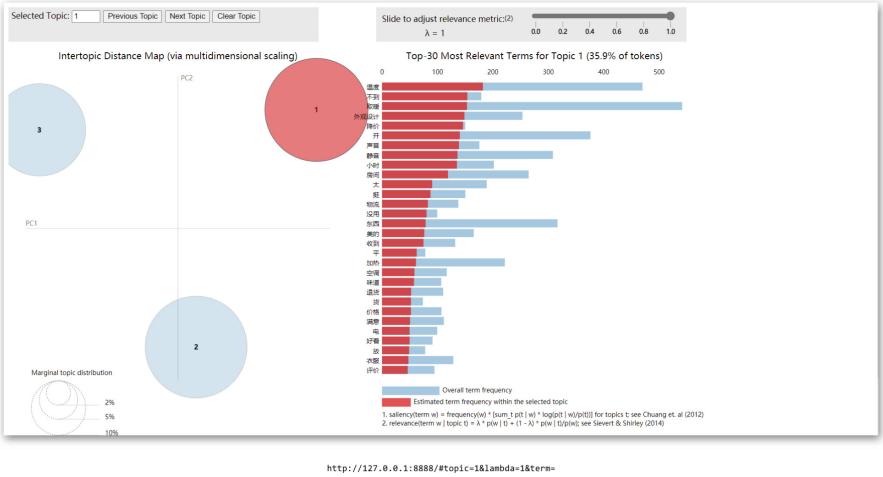

Fig. 2. Theme model diagram

4 Conclusion

Through the analysis of commodity evaluation, it can be concluded that customers attach great importance to product performance, product appearance and after-sales service. In the praise, customers mainly reflect the characteristics of fast heating, high cost

performance and low power consumption. The evaluation mainly reflects the problems that the radiator is not too hot, the heating speed is slow and the product packaging. Bad reviews mainly reflect quality problems, service problems and after-sales problems. It can be seen that the business should improve its own service problems, after-sales problems and product packaging problems. First of all, the product should improve its own performance, so that the product itself has strong performance, and customers value the cost performance and brand effect. Secondly, the product packaging should be strengthened to improve the online after-sales problem, which can respond to the product problems reflected by customers at the first time and bring better experience.

In this paper, online comments are analyzed in the text processing and analysis stages, and it is found that online comments contain hidden information that can be used. It provides a method for many e-commerce merchants to analyze online comments and provides reference for them to process online comments and mine information.

References

1. Information on http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202102/t20210203_71361.htm.
2. X.Y. Ku, R.L. Yang, L.H. Dong, Emotion Analysis of Chinese Online Commodity Evalua-tion Based on Sword2vect, Journal of Xi'an University of Science and Technology, Vol. 40 (2020) No.3, p. 504-511.
3. G.R. Zhang, C. Bao, X.Y. Wang, D.X. Gu, X.J. Yang, Text Semantic Mining and Senti-ment Analysis Based on Comment Data, Information science, Vol. 39 (2021) No.5, p. 53-61.
4. Y. Kang, H.Z. Xue, B. Hua, Evaluation and analysis of fine-grained goods oriented to deep learning network, Computer engineering and application, Vol. 57 (2021) No.11, p. 140-147.
5. J. Cao, Y.D. Zhang, J.T. Li, S. Tang, An adaptive optimal LDA model selection method based on density, Chinese Journal of Computers, Vol. (2008) No.10, p. 1780–1787.