LDA Topic Mining of Customer Text Sentiment Analysis on the E-Commerce Platform

Jiahao Duan(✉)

Safety Science and Emergency Management School, Wuhan University of Technology, Wuhan, China
duanjiahao@whut.edu.cn

Abstract. Collect the reviews left by users on the e-commerce platform for processing and analysis to understand users' needs, opinions, and product strengths and weaknesses. The text review data on the Jingdong platform was collected through a Python crawler, which was pre-processed by word separation, lexical annotation, and deactivation, and the key information of the reviews was extracted using the LDA topic model for research and analysis. The analysis of the integrated theme and its high-frequency feature words will help to obtain the valuable content and emotional orientation of the text review data and derive the advantages and disadvantages of the product and the corresponding needs of the users, providing consumers, merchants, and regulators with references and suggestions for collection.

Keywords: Text Data Mining · Sentiment Analysis · LDA Models · User Reviews

1 Introduction

With the development of science and technology and the construction of the Internet system, more and more users are choosing online e-commerce platforms for shopping with the help of big data and Internet of Things technology. E-commerce platforms are an important way to promote effective matchmaking between small and medium-sized farmers and consumers, which is conducive to ensuring the supply of fresh urban agricultural products, increasing farmers' income, and helping rural revitalization. With the support of national policies and the guidance of market winds, the agricultural e-commerce industry is developing rapidly, and the market transaction scale of China's fresh produce e-commerce industry reached 279.62 billion yuan in 2019, with an average annual growth rate of more than 50% [4]. The huge market prospect of the fresh produce e-commerce industry has attracted many large e-commerce platforms to enter the market, with famous e-commerce enterprises such as Jingdong, Alibaba, and Suning moving into this area.

Fresh produce is a necessity for people's lives, with characteristics such as short shelf life and perishability [8], and consumers have high demands on its quality. In the context of non-contact shopping, consumers' ability to obtain information about products

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 38–46, 2023.
https://doi.org/10.2991/978-94-6463-010-7_6

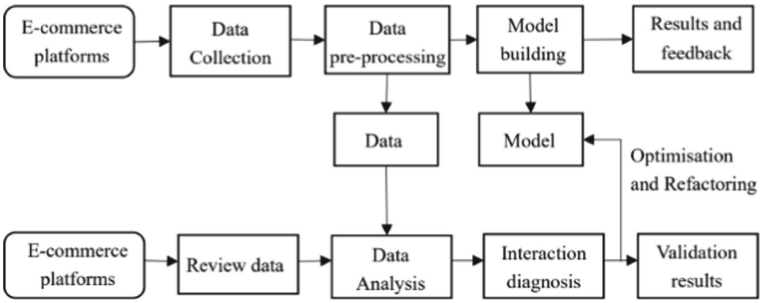

Fig. 1. User sentiment analysis process for e-commerce platform review

through personal experience is weakened, and they have to rely more on information search to obtain information that can help them make purchase decisions. However, most e-commerce platforms in the market are too rough in categorizing user reviews [11] and extracting keywords, lacking guidance value and not conducive to extracting effective information from them.

Text reviews contain information about the evaluation experience released by purchased consumers, and data mining analysis of the reviewers' text data for intrinsic information can provide consumers with purchase suggestions, suggest improvements to merchants, and provide the platform with a regulatory basis. Against this background, the author chooses to take fresh produce on the Jingdong self-operated platform as an example and conducts user sentiment analysis on user review data, to provide reference and suggestions for consumers, merchants, regulators, and other groups.

In this paper, the general process of user sentiment analysis based on review data from e-commerce platforms is studied from three aspects: data acquisition, processing, and analysis, and the process is shown in Fig. 1.

2 Literature Review

2.1 Acquisition, Pre-processing, and Representation of Online Review Texts

2.1.1 Online Review Data Acquisition

An adequate number of online reviews is the basis for conducting review sentiment analysis and impact studies, and most researchers choose comprehensive online sales platforms such as Jingdong or Taobao. Online reviews are usually obtained using both self-written crawlers [6] programs and publicly released crawler [6] software, both of which have their advantages and disadvantages.

2.1.2 Online Comment Pre-processing

After obtaining the online comment data, the garbled codes and repeatedly collected comments were first manually removed from the data collection process, and then pre-processed with word separation, deactivation, and lexical annotation.

Chinese word separation is a critical but difficult process in text analysis. Since the 1980s there have been three main types of Chinese word separation methods, namely rule-based, statistical and semantic-based [13]. Chinese lexical annotation is currently used in two ways: statistical-based lexical annotation and deep learning-based lexical annotation [14]. To train a deep learning model with good results, a large amount of annotated corpus data is required, but Wang Lianxi, Zhong Jun et al. argue that the annotated corpus of many studies is not large enough at present [12]. Deactivation refers to the automatic filtering out of words that occur too frequently and have no real meaning during the processing of data for efficiency [2]. Gong Qin and Deng Sanhong et al. proposed in 2017 that four general deactivation word lists are currently used in the research of online reviews, namely: Chinese deactivation word list, HIT deactivation word list, Baidu deactivation word list, and Sichuan University Machine Intelligence Laboratory deactivation word bank [7].

2.2 Comment Text Sentiment Analysis

Text sentiment analysis refers to the process of collecting, processing, analyzing, generalizing, and reasoning about subjective texts with emotional overtones, and involves several research fields such as artificial intelligence, machine learning, and data mining processing. Researchers often use lexicon-based sentiment analysis and LDA topic-based sentiment analysis.

2.2.1 Sentiment Analysis Based on Dictionary

The lexicon-based sentiment analysis method matches words in the text by constructing a sentiment lexicon, calculating the sentiment tendency of the comparison words, and finding the sentiment class of the whole text by pre-set rules, and the analysis process is shown in Fig. 1. The prerequisite for the method to be determined accurately is to ensure that a high-quality sentiment lexicon is available, and there are two methods of building sentiment lexicons, manual construction, and automatic construction, as applied by Gao Hualing et al. in their analysis of hotel evaluations [1].

2.2.2 Sentiment Analysis Based on the LDA Topic Model

The LDA model was proposed in 2003 by Blei, Wu et al. 2003 as an unsupervised machine learning technique [10]. Many scholars have introduced supervised mechanisms to improve the LDA model or use the LDA topic mining results as an intermediate layer to combine with other models for sentiment analysis of online reviews of fresh produce. Guo Xianda et al. [3] proposed a Gaussian LDA online review topic mining method, which applies the AP clustering algorithm to first cluster online reviews before achieving topic discovery.

3 Data Acquisition and Data Processing

3.1 Collection of Text Comments

After a preliminary market survey, it was found that jelly oranges have a wide audience among various fresh agricultural products, high nutritional value, suitable for both young

and old, easy to store and transport, ideal for online sales, and have huge consumption whether for home consumption or as a gift. Therefore, the reviews of the Jelly Orange shop in Jingdong Mall were selected as the text review collection samples. In this paper, we used Python to write a crawler program to collect the review data of a brand of apple customers from the website of Jingdong Mall. Four fields were collected: rating star, review content, product attributes, and review type, and the data was saved to an Excel table.

3.2 Review Pre-processing

3.2.1 Sentiment Analysis Based on Dictionary

Data cleaning is essential in the whole process of data analysis [13], and the quality of the cleaning results is directly related to the model effect and the final experimental conclusion. In practice, people are likely to be sloppy when writing comments and uploading them directly without careful review. Manual observation of the crawled data also reveals that there are a lot of numbers, letters, and symbols in the comments, which are of no practical value for text sentiment analysis. In addition, as the crawled data came from the Jelly Orange shop in Jingdong Mall, the text of the comments often contained many words such as "Jingdong" and "orange", which are very frequent but not useful for text sentiment analysis, and needed to be cleaned in the pre-processing stage to remove these words need to be removed in the pre-processing stage.

3.2.2 Sentiment Analysis Based on Dictionary

According to people's habits regarding product reviews, default reviews, copying others' content as reviews, are often used. This type of review is only relevant to the study if it is the first to appear [15]. Similarly, some reviews have a high degree of similarity in content, with only some differences in individual words, but removing such reviews with extremely similar text may result in a large number of mis-deletions, ultimately resulting in insufficient data. Therefore, in this study, to retain a more valid corpus that can be studied, the author only deleted completely duplicate comments and kept the information of other textual comments.

3.3 Data Processing

3.3.1 Comment Word Separation and Lexical Annotation

According to people's habits regarding product reviews, default reviews, copying others' content as reviews, are often used. This type of review is only relevant to the study if it is the first to appear [9]. Similarly, some reviews have a high degree of similarity in content, with only some differences in individual words, but removing such reviews with extremely similar text may result in a large number of mis-deletions, ultimately resulting in insufficient data. Therefore, in this study, to retain a more valid corpus that can be studied, the author only deleted completely duplicate comments and kept the information of other textual comments.

3.3.2 Extracting Reviews Containing Nouns

The main objective of this study was to obtain the product's strengths and weaknesses and the corresponding user needs and to make corresponding suggestions. Therefore, many adjectival words, such as "great", expressed the emotional inclination towards the product, but it was not possible to extract the product's strengths and weaknesses and user needs satisfaction from this textual information. In this study, only reviews with a clear noun lexical category are of value, so reviews with a noun lexical category are extracted according to lexical annotation.

3.4 Analysis of the Sentiment of Text Reviews

3.4.1 Matching Emotional Words

In this study, the main purpose of the study is to analyze the strengths and weaknesses of the product and user satisfaction, so it is not necessary to analyze the specific score of each review, as long as the emotional orientation of the review information is determined, so the lexical matching of the sentiment word matching method is used. The word list used for sentiment analysis is the "Sentiment Analysis Word Collection" from the Internet. The positive and negative word lists were combined separately and the positive words were given a weight of 1. The words in the positive word list were given a weight of 1 as the positive review sentiment word list, and the words in the negative word list were similarly given a weight of −1 as the negative review sentiment word list. The researcher optimized the word list for the characteristics of the online reviews and the goods in this study.

3.4.2 Modifying Sentiment Tendencies

In text reviews, many users add negative words before positive words to express the opposite meaning, and even in Chinese, there is the phenomenon of multiple negations, i.e., when the negative word appears again an odd number of times to indicate negation, and an even number of times to indicate affirmation, so it is important to make corrections for sentiment tendency. The correction of affective tendency is generally based on the lexicality of the two words preceding the affective word to determine whether the effective value is positive or negative. After reading in the negative word list, the above results were corrected for sentiment value tendency, the sentiment score of each comment was recalculated, the comments were divided into positive and negative ratings, and finally, the text sentiment analysis accuracy rate was calculated. From the experimental data, the accuracy rate of the sentiment analysis based on the word list reached 86.93%, which shows that the text sentiment analysis is effective.

4 Theme Analysis Models Based on LDA Models

4.1 Optimal Number of Topics

In the data processing part, two datasets of positive and negative reviews after text sentiment analysis are obtained according to the ratings of text sentiment ratings and

using these two datasets, the author builds the dictionary and corpus. In the optimal LDA model selection method, the number of topics is determined by calculating the cosine similarity between the topics to derive and measure the degree of similarity, the more similar the meanings of the words used, the more similar the content.

In this study, the author wanted to find out the keywords under different topics by using the LDA topic model. Therefore, in this step, the top 100 words in each model are taken out and combined into a set to generate a vector of word frequencies between any two topic words, and the cosine similarity is calculated according to the above formula, with larger values indicating greater similarity. By calculating and visualizing the average cosine similarity of each topic number, it can be obtained that the most appropriate topic to select for positive and negative evaluation into the LDA model is 3. The results are shown in Table 1, Table 2, and Fig. 2.

Table 1. The average similarity of positive review topics

1	2	3	4	5	6	7	8	9	10
1	0	0	0.01	0.008	0.007	0.006	0.008	0.006	0.002

Table 2. The average similarity of negative review topics

1	2	3	4	5	6	7	8	9	10
1	0.34	0.02	0.05	0.03	0.07	0.05	0.09	0.104	0.135

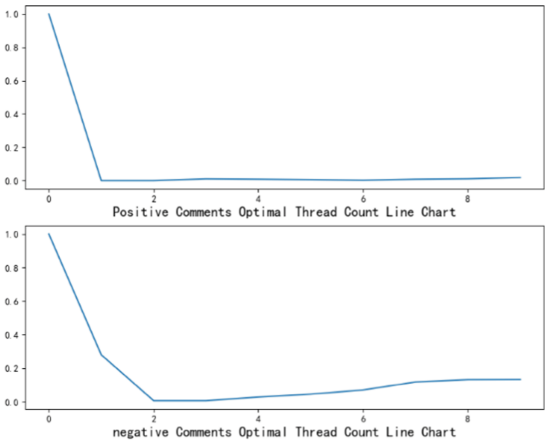

Fig. 2. Positive and negative evaluation optimal theme line chart

4.2 LDA Theme Analysis Model Results

The LDA topic analysis was completed using the Gensim library in Python based on the results of the optimal number of topics. The review texts were clustered into 3 topics and each topic generated 10 most likely words and their corresponding probabilities, the results were ranked in order of probability and represented in a table as shown in Table 3 and Table 4. Table 3 represents the topic word cloud for positive text reviews of a brand of Jelly Orange, and Table 4 represents the topic words for negative text reviews.

The above themes and high-frequency words show that the strengths of a brand of jelly oranges on Jingdong self-managed platform are high quality, good packaging,

Table 3. Positive evaluation of potential topics

Topic 1	Topic 2	Topic 3
Eat	Good food	Good
Satisfied	Fruit	Fresh
Packed	Quick	Taste
Too	Packaging	Good
Worth	Shipping	Not bad
Cheap	Fruit	Taste
Activity	Cheap	Quality
Very good	Good	Okay
Like	Oranges	Little
Buy	Two boxes of	Peels

Table 4. Negative evaluation of potential topics

Topic 1	Topic 2	Topic 3
Bad	High	Rubbish
Expensive	Quality	Eat
Fruit	Rotten	Fresh
Price	Peel	Expensive
Buy	Poor	Taste
Pit	Disappointed	Eat
Really	Sour	Bad
Orange	Fruit	Jelly
Poor	Shopping	Self-employed
Eat	Difficult to eat	Size

high-cost performance, and fast shipping; while the dissatisfaction of users lies in poor product quality control, large differences between batches and relatively high prices.

Through the LDA thematic model analysis of the user evaluation of a brand of jelly oranges on the Jingdong self-operated platform, the author makes the following suggestions to the brand's shop: ① Under the premise of ensuring good packaging and high-cost performance of jelly oranges, stabilize the quality of each batch of products, strictly control in the sorting of jelly oranges, so as not to use substandard as good or small as big, and improve the freshness of the products, so as not to affect the reputation of the brand among consumers. This will not affect the brand's reputation among consumers. ② Grading the products for consumers and increasing the types of products with smaller packaging and higher cost performance. For consumers, the author recommends that they choose to buy during the peak fruit season so that they can buy fresher batches of jelly oranges with better quality products. At the same time, if there is dissatisfaction with the product, to promptly contact for exchange processing. At the same time, as Jingdong's self-managed platform of goods, as a regulatory platform of Jingdong Mall to be responsible for the products sold on its platform, to provide better transport, packaging, after-sales service, cannot fail to consumer trust.

5 Conclusion

This paper shows the complete process of data collection, pre-processing, analysis, modeling, and evaluation of user reviews of a product on an e-commerce platform. After the steps of cleaning, de-duplication, word separation, and extraction of noun comments on the review data, the model of text sentiment tendency analysis and LDA topic analysis was used to analyze the user review data for text sentiment analysis, and its accuracy rate reached 86.93%, which is a good result. In this study, the author took a brand of fresh produce jelly orange from the Jingdong self-operated platform as an example, analyzed user reviews, and the results of the study were used as a basis to propose rationalized sales, purchase, and management suggestions to merchants, consumers, and regulatory platforms in turn.

Due to the author's conditions, the data collection for this experiment is small, the text topic clustering effect is not ideal, the topic analysis is not perfect, and the LDA model results are not good enough. Subsequent research can be optimized in the following aspects: ① Increase the amount of data collection, further optimize the model and improve the evaluation effect. ② Conduct comment text sentiment analysis from multiple aspects and industries to achieve more accurate sentiment analysis of Chinese text comments.

Acknowledgments. I am grateful to Li Shixuan and Liao Binzhou for their useful guidance and selfless help. Also, I would like to thank all the friends who are always there. I should finally like to express my gratitude to my beloved parents who have always been supporting me without a word of complaint.

References

1. Gao H, Zhang J (2021) Sentiment analysis and visualization of hotel reviews based on sentiment dictionary. Software 42(01):45–47+66
2. Guan Q, Deng S, Wang H (2017) A comparative study of common stop word lists for Chinese text clustering. Data Anal Knowl Disc 1(03):72–80
3. Guoxianda N, Gao H, Yang X (2020) Research on online review topic mining based on Gaussian LDA. J Intell 39(06):630–639
4. Hu Y, Lin H (2021) Influence of online review features on fresh e-commerce produce sales - evidence from Taobao lamb big data. J China Agric Univ 26(06):206–218
5. Li B (2021) Coping with the first "no late night" Jingdong Double Eleven. China Storage Transp 12:58
6. Li Y (2021) Research on web crawler technology based on Python. Electron World 03:39–40
7. Lv J (2020) Research on online hotel reviews based on text mining. Huazhong Normal University
8. Ma R (2021) Research on the purchase intention of fresh produce e-commerce consumers. Mod Mark (Lower J) 02:66–67
9. Tian Y, Gong T (2021) Research on theme mining of online teaching demand data based on the LDA model. Intell Sci 39(09):110
10. Wang D, Zhao R, Kou Y, Xian G (2017) LDA model-based findings on the research themes of agricultural research projects in the last 10 years of the EU Framework Programme. Agric Outlook 13(04):69–75
11. Wang X, Sun Y (2002) Quality signals in China's food market. China Rural Econ 05:27–32
12. Xu Y, Zhu S (2020) A statistical study on the annotation of verb-name and class words in HSK vocabulary level syllabus. J Yunnan Normal Univ (Foreign Chin Teach Res Ed) 18(04):51–57
13. Yang G, Xu X, Feng L, Xu Y (2019) A likelihood-oriented Chinese word separation method based on a maximum matching algorithm. Stat Inf Forum 34(03):18–23
14. Zhou T, Mo L, Hu M, Li V (2020) A Chinese lexical annotation method based on MEM and HMM. Journal of Jishou Univ (Nat Sci Ed) 41(02):15–18
15. Zhou Y, Bai J (2020) Research on sentiment analysis of e-commerce reviews. Small Medium-Sized Enterp Manag Sci Technol (Zhongjun J) 06:130–131