International Scientific Technical and Economic Research | **ISSN: 2959-1309 | Vol.3, No.2, 2025**
www.istaer.online——Research Article

Comment Analysis of Online Shopping Based on Big Data and NLP Techniques

Dongping Sheng, Zhongyuan Ma, Haidong Feng, Chenqi Zhou, Hao Liu, Hun Guo, Chun Su*

Changzhou Institute of Technology, Changzhou, China

Received: 11 March 2025
Revised: 23 March 2025
Accepted: 31 March 2025
Published: 1 April 2025

Abstract: With the rapid development of online shopping, the number of consumers has increased significantly, and user reviews have become increasingly influential on sellers and brands. User reviews not only provide feedback on products and services, but also provide companies with important market insights. Therefore, review analysis has become a crucial research field. With the help of big data, artificial intelligence (AI) and natural language processing (NLP) technologies (such as keyword extraction, sentiment analysis, etc.), valuable information can be effectively extracted from massive consumer reviews. To this end, we designed a mathematical model based on ASIN (Amazon Standard Identification Number) for in-depth analysis of product review text. Through ASIN, relevant data on the Amazon website, including product names, categories and other information, are obtained, and the data is cleaned, classified and processed. Finally, we generate word clouds and construct relationship network diagrams to show the potential patterns and connections in text data, providing data support and visual analysis for product and market decisions.

Keywords: Word cloud map; VADER model; Review sentiment analysis model; MCDS model

1 INTRODUCTION

The rise of online shopping has led to an abundance of customer reviews on e-commerce platforms. Positive reviews can significantly boost purchases, while negative reviews can deter a large number of potential buyers. Consumers often face information overload and decision - making challenges due to the vast amount of product information and choices available [1]. Through NLP (Natural Language Processing) based on deep learning, keyword extraction, opinion mining, and emotional analysis can be performed on customer feedback data, leading to intuitive visualization and digital enhancement of consumer insights [2]. By making full use of these data, it can achieve a win - win situation for both the buyers and the platform.

For example, for the buyers, they can quickly understand the pros and cons of products through the visualized consumer insights [3-6]. They no longer need to spend a lot of time screening a large amount of text information. Instead, they can directly see key points such as the most praised features and the most complained problems of the products. This helps them make more informed purchasing decisions [7].

For the e-commerce platforms, they can use these analyzed data to improve product recommendations. By understanding the real needs and concerns of consumers, platforms can

recommend products that are more in line with users' interests. Moreover, platforms can also use these data to supervise merchants. If a certain merchant has a large number of negative reviews in a certain aspect, the platform can take measures to urge the merchant to improve, which helps to maintain the overall quality of products on the platform and enhance the platform's competitiveness [8]. In addition, these data can also be used for market research, helping platforms to predict market trends and adjust their business strategies in a timely manner.

2 ASSUMPTIONS

(1) In a comment text, stop words such as "and" and "or" can be ignored during the analysis of the comment text.

(2) A piece of text can be represented by several keywords, and by quantifying and combining these keywords, it can be used to represent a product review.

(3) The rating of a product is determined by the emotional keywords in the product reviews, and the mapping relationship between the reviews and the rating can be found by using emotional keywords.

3 MODEL ESTABLISHING AND ANALYZING

3.1 Model establishing

Before data analysis, it is necessary to ensure the validity of the data, so before building the model, it first processes the data provided in the appendix, including data cleaning, data classification, text pre-processing and so on [9].

Step 1: Data cleaning

Observing the data in the given appendix, we found various information such as comment number, rating result, and comment content placed in a single column of the Excel table. To facilitate subsequent analysis and model building, the data was disaggregated by using Excel functions. Afterward, the team checked for missing values and outliers, removing the irrelevant "timestamp" column. (Refer to Table 1 for processing results.)

Table 1: GA-BP neural network parameter settings

Pretreatment			
{"reviewer ID": "A11N155CW1UV02", "asin": "B000H00VBQ", "reviewerName": "Adriana M", "helpful": [0, 0], "review Text": "I had big expectations because I love English TV, in particular Investigative and detective stuff but this guy is really boring. It didn't appeal to me at all.", "overall": 2.0, "summary": "A little bit boring for me", "unix Review Time": 1399075200, "review Time": "05 3, 2014"}			
After treatment			
Reviewer ID	asin	Reviewer Name	helpful
A11N155CW1UV02	B000H00VBQ	Adriana M	[0, 0]
Review Text	overall	summary	Review Time
I had big…	2.0	A little…	05 3, 2014

Step 2: Data crawling

Based on the datasets' characteristics, it can be found that the appendix contains publicly available Amazon datasets from 2013 and 2014, with ASIN as the unique product identification code on Amazon. To ensure data analysis accuracy and completeness, and to facilitate model testing, it crawled product names, categories, and ratings corresponding to the reviews using "ASIN" as the classification and testing datasets.

Step 3: Data classification

By connecting the cleaned datasets from Step 1 with the crawled datasets from Step 2, aligning related data to obtain richer, more comprehensive, and relevant information. This process provided valuable insights for subsequent data analysis and decision-making [10]. Using Python, the data was classified based on product categories and ratings, resulting in multiple sub-datasets under each appendix. Each sub-dataset corresponded to a specific product category or rating, simplifying the subsequent analysis and comparison.

Step 4: Text preprocessing

(1) Remove special characters: First, remove special characters in the text, such as punctuation marks, numbers, symbols, etc., to ensure that only plain text is retained.

(2) Segmentation: Segment the text into words or phrases to form a vocabulary list.

(3) Deactivation: Stop words are common words that have no practical meaning in text analysis, such as 'the', 'and', 'is', etc. These words need to be removed from the text. These stop words need to be removed from the text to reduce data noise.

(4) Stemming / Word Form Reduction: The vocabulary is stemmed, i.e., the various forms of a word are reduced to its original stem form. This helps to further reduce the redundancy of words and improve the efficiency of text analysis.

Step5: Keyword extraction

Finally, filter out some irrelevant or unrepresentative keywords according to the set keyword filtering conditions, to get the final keyword list, and then perform word frequency statistics and visualization analysis according to overall and product categories.

3.2 Model solving

(1) Word frequency statistics and analysis

Through data processing and text analysis, it can get the overall noun and adjective word frequency of Appendix I as shown in Figure 1:

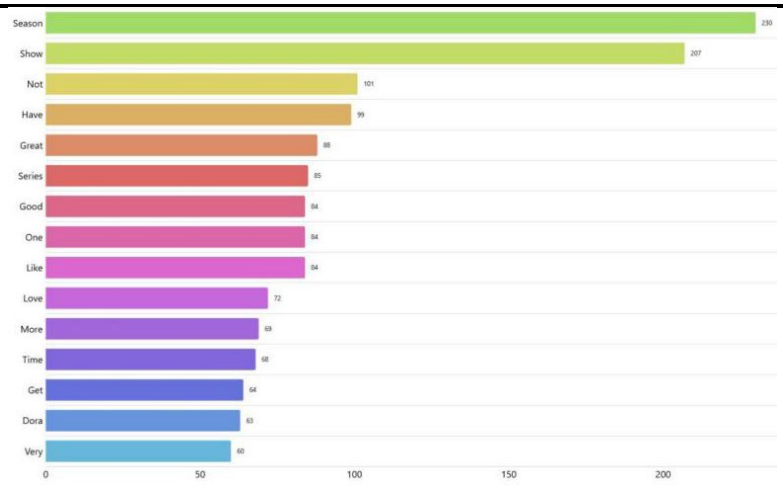

Fig. 1: Appendix I Word Frequency Statistics Chart

From the noun word - frequency statistics chart, it can be roughly concluded that the goods in Appendix I are in the category of movie and television dramas, which involve romance and interludes. From the adjective word - frequency statistics graph, it can be concluded that the overall rating of the audience is favorable, and it can be concluded that the quality of the movie and television dramas in Appendix II meets the public's expectations.

(2) Word cloud maps

Each product is rated from 1-5, and ratings 4 and 5 are set as grade A, rating 3 as grade B, and ratings 1 and 2 as grade C. Next, word clouds are plotted separately for each rating, as shown below:

Fig. 2: Appendix I Word cloud map

Figure 2 shows that the core keyword in Appendix I is 'show', indicating it relates to film and television content. 'Season' is prominent for film and television dramas, and other terms like 'variety', 'full', etc., provide descriptions and evaluations. In Figure 2, many sentiment words in rank B product reviews contain both positive and negative sentiments, like 'bad and good.' This aligns with the expected rating grade, reflecting the accuracy of our analysis.

4 CONCLUSIONS

The analysis of online shopping reviews using big data and natural language processing (NLP) techniques has provided significant insights into consumer behavior, product sentiment, and market trends. The designed mathematical model, based on the ASIN (Amazon Standard Identification Number), effectively captured and analyzed the product reviews from e-commerce platforms. By cleaning, classifying, and processing the data, we were able to extract valuable keywords and perform sentiment analysis. The generated word clouds and

relationship network diagrams highlighted key product features, common customer concerns, and overall product ratings.

The results from the sentiment analysis showed that positive reviews were correlated with favorable product attributes, while negative reviews often pointed to issues that could be improved. The word cloud maps further confirmed this, as they visually represented the most frequent terms associated with high and low ratings. For instance, products with ratings 4 and 5 were associated with positive words like "good" and "great," while products rated 1 and 2 showed more frequent use of words like "bad" and "disappointing."

5 DISCUSSIONS

This study demonstrates the power of integrating big data, AI, and NLP techniques for online review analysis. The approach not only enables efficient keyword extraction and sentiment analysis but also provides actionable insights for both consumers and e-commerce platforms.

For Consumers: The visualized insights, such as word clouds, provide an intuitive and quick understanding of the key features of a product. Consumers can easily identify the strengths and weaknesses of a product, enabling them to make more informed purchasing decisions without having to go through extensive text-based reviews. The ability to quickly assess what others like and dislike about a product simplifies the decision-making process and reduces the time spent evaluating products.

For E-commerce Platforms: Platforms can leverage the findings from review analysis to improve product recommendations and tailor their offerings to better match consumer interests and preferences. Moreover, sentiment analysis allows platforms to monitor merchant performance. A sudden increase in negative reviews in a specific product category could trigger a platform's alert system, prompting action to ensure quality control and maintain a positive user experience. By understanding common consumer complaints, platforms can also provide merchants with detailed feedback to address potential issues, improving both product quality and customer satisfaction.

Additionally, the review analysis can assist platforms in market research, helping them predict emerging trends and adjust their strategies accordingly. For instance, by analyzing the sentiment and keywords associated with certain product categories, platforms can identify the next popular product or detect shifts in consumer preferences, giving them a competitive edge in the market.

Challenges and Future Work: While this study presents a promising approach, several challenges remain. First, the complexity of human language poses difficulties in perfectly interpreting sentiment, as opinions can often be ambiguous or mixed. Moreover, NLP techniques are not always able to capture the full context of a review, potentially leading to misinterpretation of sentiment. Future work could focus on refining sentiment analysis models, integrating more advanced techniques like deep learning, and incorporating contextual information to enhance the accuracy of sentiment detection.

In addition, this model could be expanded to include reviews from multiple platforms, not just Amazon, to provide a broader view of consumer opinions across different e-commerce sites. This would further enhance the robustness of the model and improve its applicability in various market scenarios.

ACKNOWLEDGEMENTS

This work is supported by ministry of education industry-university cooperative education project (Grant No.: 231106441092432), the research and practice of integrating

"curriculum thought and politics" into the whole process of graduation design of Mechanical engineering major: (Grant. No.: 30120300100-23-yb-jgkt03), research on the integration mechanism of "course-training-competition-creation-production" for innovation and entrepreneurship of mechanical engineering majors in applied local universities (Grant. No.: CXKT202405), Mechanical manufacturing equipment design school-level "gold class" construction project (Grant. No.: 30120324001).

REFERENCES

[1] Zhang Z. Q., & Ye Q., (2010). A Review of Sentiment Analysis Research on Intern et Product Reviews. *Journal of Management Science*, 13(006): 84-96. DOI: https://doi.org/10.54097/fbem.v10i1.1017

[2] Chang, W., & Zhu, M. (2023). Sentiment analysis method of consumer comment te xt based on BERT and hierarchical attention in e-commerce big data environmen t. *Journal of Intelligent Systems*, *32*(1), 20230025. DOI: https://doi.org/10.1515/jisys-2023-0025

[3] Mars, A., & Gouider, M. S. (2017). Big data analysis to features opinions extractio n of customer. *Procedia computer science*, *112*, 906-916. DOI: https://doi.org/10.1016/j.procs.2017.08.114

[4] Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for re commending shoppers, shops based on customers' review. *Complex & Intelligent Sys tems*, *6*(3), 621-634. DOI: https://doi.org/10.1007/s40747-020-00155-2

[5] Wang, J., Shu, T., Zhao, W., & Zhou, J. (2022). Research on Chinese consumers' at titudes analysis of big-data driven price discrimination based on machine learnin g. *Frontiers in Psychology*, *12*, 803212. DOI: https://doi.org/10.3389/fpsyg.2021.803212

[6] Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand's s ocial media marketing on customer engagement: Using big data analytics and nat ural language processing. *Journal of Business research*, *125*, 815-826. DOI: https://doi.org/10.1016/j.jbusres.2019.04.042

[7] Ranjan, M., Tiwari, S., Sattar, A. M., & Tatkar, N. S. (2024). A New Approach for Carrying Out Sentiment Analysis of Social Media Comments Using Natural Langu age Processing. *Engineering Proceedings*, *59*(1), 181. DOI: https://doi.org/10.3390/engproc2023059181

[8] Tian, J., Zhang, J., Han, K., Qiao, J., & Li, P. (2023, November). Power Customer Satisfaction Based on Power Big Data and NLP. In *International Conference on Cogni tive based Information Processing and Applications* (pp. 377-387). Singapore: Springer N ature Singapore. DOI: https://doi.org/10.1007/978-981-97-1975-4_34

[9] Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., & Yan, X. (2018). Measuring cu stomer agility from online reviews using big data text analytics. *Journal of Manage ment Information Systems*, *35*(2), 510-539. DOI: https://doi.org/10.1080/07421222.2018.1451956

[10] Aldabbas, H., Bajahzar, A., Alruily, M., Qureshi, A. A., Amir Latif, R. M., & Farh an, M. (2020). Google play content scraping and knowledge engineering using nat ural language processing techniques with the analysis of user reviews. *Journal of In telligent Systems*, *30*(1), 192-208. DOI: https://doi.org/10.1515/jisys-2019-0197