Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230934

Improved TF-IDF-Based LDA Topic Clustering Model for Specific Commodity Contexts and Different Sentiment Tendencies

Hao CHEN[1†], Yile ZHU[†]
College of Modern Postal Xi'an University of Posts and Telecommunications, Xi'an, China
[†]*These authors contributed equally*

Abstract. As Natural Language Processing (NLP) is increasingly used in Internet content platforms, more businesses and organizations are concentrating their research and development efforts on NLP. Effective user review data mining has the potential to greatly accelerate the transition of digital products, particularly in e-commerce. The information represented by the review texts of various shopping categories differs, but due to the "generality" of their construction models, many traditional NLP text libraries frequently make judgments about a given context that are incorrect. This error is primarily due to a lack of a precise understanding of the "characteristics" of various domains. This research chooses to review data for three different product categories (clothing, cosmetics, and laptops) with distinct contexts on the Jingdong e-commerce platform based on the aforementioned problems. To analyze and examine the classification of consumers' subject terms under different emotional tendencies for two different contexts of negative text and positive text, we propose the ED-LDA (Emotion discrimination - Latent Dirichlet Allocation) topic model. The findings demonstrate the significance of subject word grouping under various sentiment trends and context-specific characterized natural language analysis for social media review analysis of niche goods and subcultural circles.

Keywords-Text mining; sentiment analysis; SnowNLP; LDA topic model; TF-IDF

1. Introduction

1.1. Status of Natural Language Processing (NLP) Research in E-commerce

Online reviews can visually reflect the reputation and brand influence of merchants thanks to the ongoing development of e-commerce platforms, and they are an essential part of the decision-making process for customers. Online reviews fall under the umbrella of Internet Word of Mouth Marketing (IWOM), which describes customers who use e-commerce platforms to buy goods, evaluate those purchases based on how much they cost after delivery, and interact and talk to other customers in the merchant review section [1]. Online reviews have been shown to influence both consumer groups'

[1] Corresponding Author: Hao CHEN, College of Modern Postal Xi'an University of Posts and Telecommunications; e-mail: chmike666@stu.xupt.edu.cn

purchasing decisions as well as the overall sales of goods on e-commerce platforms, according to a 2007 theory put up by academicians Danny Weathers and others [2]. To succeed, businesses must comprehend the wants, feelings, and behaviors of their clients [3]. Noteworthy, sentiment analysis is in charge of solving related issues and is essential to text classification and polarity detection. Currently, the majority of e-commerce platforms' online review mechanisms are aimed at customers who have already made purchases, which means that the more reviews a product has, the better its sales are, which in turn suggests that the majority of customers support and recognize it. The cost of the transaction is lower and there is less ambiguity in the buying choice in this instance [4]. As a result, the quantity of online reviews, a key indicator of evaluation, has a big effect on product sales [5].

1.2. Current Status of Sentiment Analysis

By reviewing and sorting through the relevant literature in recent years, we find that sentiment analysis is not a brand-new topic, but its research methods are up-to-date. It can be seen that scholars have conducted a large number of studies in the field of sentiment analysis and have devoted themselves to understanding customers' [6], [7] emotions. Sentiment analysis is a technique for identifying ambiguities in language, opinions, etc. It is also known as "opinion mining" which reveals how speakers and users feel about a particular topic [8]. In addition, this paper notes that both domestic and foreign experts have conducted a significant amount of research on the topic of online review sentiment analysis. A. Moreo et al. (2012) suggested building a more thorough sentiment propensity analysis system based on a sentiment lexicon with a dedicated corpus for sentiment classification of extracted consumer opinion attitudes in product reviews and discovered that the classification effect was significantly enhanced. Shoushan Li created the first sentiment dictionary in China using references from English dictionaries, however, it also has the drawback of having limited practicality. As a result, a growing number of academics have added sentiment words from their own research disciplines to the fundamental sentiment dictionaries already in existence, creating specialized sentiment dictionaries that have increased the accuracy of sentiment categorization.

1.3. Current Status of Subject Word Extraction

Since David M. Blei and other scholars published Latent Dirichlet Allocation [9] in 2002, the academic exploration of probabilistic models for topic words in different scenarios and corpora has not stopped.

By extracting document subject keywords in LDA, we can get the subject words under different corpora and thus study the conclusions or factors obtained in a certain context. Scholars such as Marcio Pereira Basilio [10] generated information that can effectively identify local crime needs based on knowledge discovery methods in an emergency response database of police incident reports. Mingyue Zhou [11] studied a corpus of the new variety of rights dispute decisions of Chinese plants and obtained the main causes of disputes. Scholars such as Jiying Wu [12] summarized the qualities required to be a scientific and technological talent by studying the textual materials in biographies and interviews of Chinese scientific and technological talents. scholars such as Xuefeng Wang [13] constructed a model for evaluating the technological competitiveness of enterprises to provide rich information about the competitive

landscape of a certain field by studying the evaluation content of patent competitiveness. Han Gang and Li Menggang et al. constructed a transportation development model by mining the themes contained in news data and analyzed the development trends of transportation in various countries [14]. Muhammad Inaam ul Haq [15], while studying the field of the Internet of Things (IoT), used the LDA topic model to theme mine the titles and abstracts of relevant articles published between 2008 and 2020 to summarize a large number of research hotspots and trends in IoT.

Academics have simultaneously developed a large number of LDA models that incorporate advancements over existing approaches. For instance, Shaorong Feng [16] introduced the LDA model with a time-variant balancing treatment and mined the evolution of heat and topic intensity using microblogging social platforms as a corpus. In order to direct the assignment of subject terms in the LDA learning process, Changxuan Wan [17] and other researchers created Association-Constrained LDA (AC-LDA), which greatly increased the keyword extraction accuracy.

1.4. Our work

In this paper, we collected the review data of three types of products, namely, clothing, laptop, and cosmetics, from the Jingdong e-commerce platform through a Python crawler program, reserved special words, and set deactivated words according to specific rules through the Jieba library, and then performed pre-processing operations such as word separation, followed by using SnowNLP to score the sentiment of each review text, and combined with the machine learning classifier K- Means to classify the review data into negative and positive reviews. And finally, the keywords are extracted from the LDA topic model after combining the Term Frequency-Inverse Document Frequency (TF-IDF) with these classified review data, which can effectively reflect why consumers are enthusiastic about these products and why they reject them. Figure 1 shows the workflow framework.

1.5. Innovations of this paper

1.5.1 Innovations in data processing

Today's third-party technology libraries for data processing are frequently constructed using large language models, which frequently neglect the study of lexicality and word separation in specialized corpora due to their need for generalization. As a result, when used with specialized and abstract texts, these libraries' accuracy suffers significantly.

Before building the models for the three types of goods in this paper, we thoroughly and in-depth researched their histories, cultures, and Internet buzzwords. We also handled some brand names, product abbreviations, and player flirtation words very carefully, preserving the integrity of a large number of unique words and minimizing the impact of the large language model on their accuracy.

1.5.2 Innovations in technical tools

Prior to extracting the keywords, we classified the review corpus into negative and positive reviews using sentiment analysis. By creating LDA topic models with various sentiment colors, we were able to better understand the positive and negative choices that consumers make when making purchases.

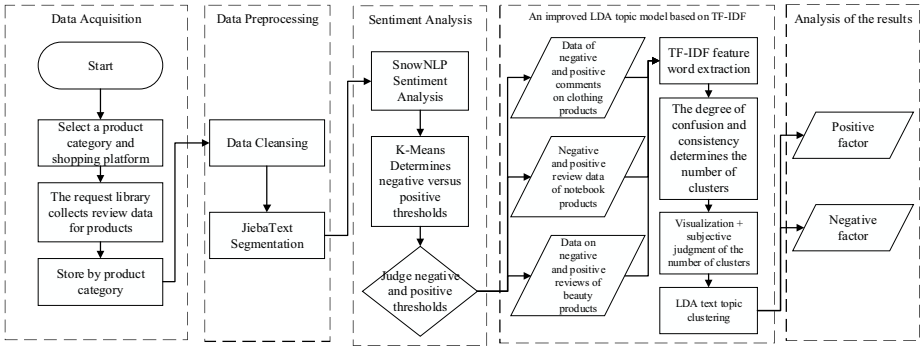

Figure 1. Framework

2. Methodology

2.1. SnowNLP

The examination of text data from social hotspots, online comments, individual opinions, and other topic qualities is known as sentiment analysis. It is a crucial component of natural language processing (NLP), and unsupervised machine learning is the main method used to study it.

The emergence of SnowNLP libraries more than makes up for the fact that traditional natural language processing libraries are mostly focused on English and have poor compatibility with Chinese text processing [18]. With a Chinese positive and negative sentiment training set included, the SnowNLP library is highly suited for processing Chinese text data. It primarily uses the basic Bayesian principle to achieve sentiment analysis, lexical annotation, text classification, and other operations.

2.2. TF-IDF

The TF-IDF algorithm, sometimes referred to as the word frequency-inverse text frequency algorithm, is a widely used weighting method for text mining and information retrieval. The basic contention is that with the exception of stop words, a word's value is inversely correlated with its frequency in the corpus and directly correlated with its frequency in the article.

TF (Term Frequency) reflects the number of times a word appears in the document it is in, Equation (1) is the calculation method of TF, where N_ω represents the number of occurrences of the word ω in a text, and N is the total number of terms in the text.

$$TF_\omega = \frac{N_\omega}{N} \tag{1}$$

IDF (Inverse Document Frequency) reflects the frequency of a word appearing in the entire document. Equation (2) is the method of calculating IDF, where Y is the total number of documents in the corpus and Y_ω is the number of documents containing the term ω.

$$IDF_{\omega} = log\frac{Y}{Y_{\omega}+1} \quad (2)$$

TF-IDF is equal to the value of TF multiplied by the value of IDF, as shown in Equation (3). A more considerable TF-IDF value for a word indicates the more significant importance of the word to the text and is easier to retain when keyword extraction of the text, while a word with a smaller TF-IDF value is more likely to be discarded.

$$TF - IDF = TF_{\omega} * ID \quad (3)$$

2.3. LDA Topic Model

LDA (Latent Dirichlet Allocation) is a document topic generation model that consists of a three-layer structure of words, topics, and documents. The word frequency of each word in each document is determined using word segmentation, and the "document-word" matrix is then created. This is the basic idea behind the LDA topic model algorithm. The document's topics are examined once the document-word matrix, "subject-word" matrix, and "document-topic" matrix are obtained through training. Each word's likelihood of appearing in the text is indicated as:

$$\mathrm{P(word|doc)} = \sum_{\mathrm{topic}} \mathrm{p(word|topic)} \times \mathrm{p(doc|topic)} \quad (4)$$

This paper improves the traditional LDA topic model based on the TF-IDF algorithm, and the specific process is as follows:

- Positive and negative data sets for the three categories of goods were obtained after sentiment classification.
- Invoke gensim library, combine TF-IDF feature word extraction and corpora module to construct dictionary and corpus.
- Use model. LdaModel() specifies the number of topic topics for LDA model training.

3. Experiments

3.1. Data Acquisition

At this stage, JD.com has become the largest 3C online shopping platform in the domestic B2C market. As a representative of domestic e-commerce shopping platforms, the business development of JD.com has received extensive attention. In order to compare the data of different shopping categories on JD.com, this paper selects suitable themes to better represent the main characteristics of different shopping categories. Observing the best-selling hot lists of JD.com over the years, it can be found that laptops, clothing, cosmetics, and other products are its best-selling categories. Therefore, for the three shopping categories of laptops, clothing, and cosmetics, considering the best-selling and representative degree of products, as well as the influence of seasonality, this paper selects the top five best-selling theme products under the corresponding categories as the analysis objects for product review data acquisition.

The reviews of these products are effectively crawled in this study using crawler technology, which takes into account the legitimacy and volume of the reviews while also selecting the first 30 pages of text content comments that fall within its recommended ranking for crawling and the JSON file of the JD product page for selecting the review content field. Use Python tools and the Requests module in particular; the crawl time is January 11, 2023.

3.2. Data Preprocessing.

It is required to perform data preprocessing on the original text corpus after acquiring the original comment text corpus in order to mine and analyze the intrinsic information of the original corpus. Data cleaning and Chinese word segmentation are two examples of the data preprocessing processes used in this study.

3.2.1 Data Cleansing

Data cleaning is done to get rid of redundant information, fix mistakes that are already there, and ensure consistency. In this study, the data cleansing work mostly consists of:

- Delete system labels, such as “Appearance,” “Running Velocity,” and “Package Protection.” The system automatically generates system labels in user comments. They are neutral words without emotional bias. At the same time, they will interfere with the science and authenticity of following word frequency statistics, so this paper deletes system labels.
- Delete meaningless Spaces and commodity names, such as “Armani,” “Nike,” and “Y7000P.” Meaningless Spaces and commodity names do not have emotional bias and seriously interfere with word frequency statistics, so this paper also deleted them. It is worth noting that in the category of laptops, after careful observation, it is found that “Samsung, Micron, and BOE” are also brand names. However, they have no direct relationship with the product brand belonging to the hardware brand of the commodity, moreover; the appearance of such brands often means that consumers praise that the product uses hardware from a prominent manufacturer, which is commendable in itself, so it is reserved here.
- Eliminate emojis. The emoji symbols on JD.com consist of & plus a word. In order to prevent ambiguity caused by the difference between the meaning of the word itself and the emotional tendency of the comments, this paper removes the emoji symbols.

3.2.2 Comment on Data Segmentation

In text mining, word segmentation is the initial stage and a crucial component of data preprocessing. The first stage in this process is to separate the original text into the smallest possible chunk of knowledge that still makes analytical sense. The Jieba package for Python is used in this article to segment words. In order to create word segmentation results for the creation of phrases with a high chance of association between Chinese characters, Jieba, a great third-party library for Chinese word segmentation, uses a Chinese thesaurus to determine the probability of association between Chinese characters. It is important to note that the thesaurus is disabled for blocking in this topic and is only specified to be retained in the laptop category. The proprietary thesaurus (e.g., "RTX3060," "YYDS") is one of them, and its purpose is to

maintain the integrity of the product-specific words in the review in order to prevent them from being dropped during word segmentation and losing crucial information. Word segmentation's disruptive effects are blocked when the thesaurus is disabled.

3.3. Sentiment Analysis

This study uses SnowNLP to analyze the sentiment of the preprocessed data and separate the comment text into two categories: positive and negative evaluations.

3.3.1 SnowNLP Sentiment Score Calculation

Using the SnowNLP module, the sentiment score of the comments after data cleansing is calculated, and the results obtained by the visualization output are shown in Figures 2, 3, and 4.

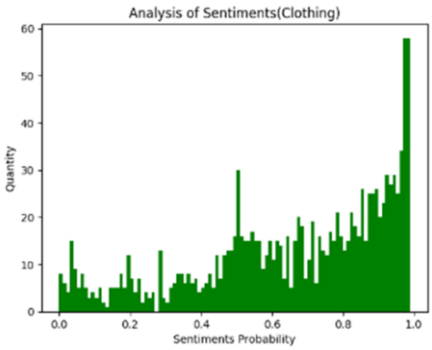

Figure 2. Clothing

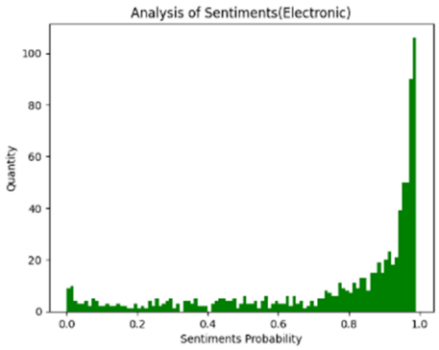

Figure 3. Laptops

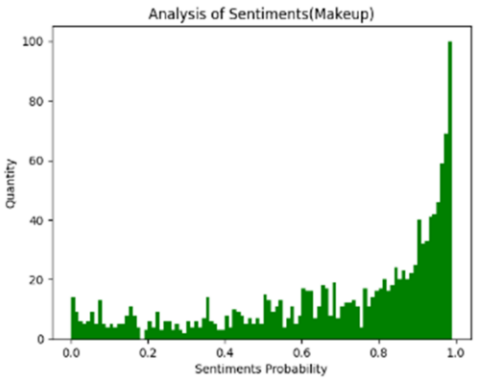

Figure 4. Cosmetics

3.3.2 K-Means data classification

The sentiment score should fall between 0 and 1, with a closer score of 0 indicating a worse mood and a closer score of 1 indicating a higher likelihood of positive emotions [19]. It is important to note that according to Xia Yuqin and Shan Xuewei scholars, a positive feeling is one that is larger than 0.5 and a negative emotion is one that is less

than 0.5 [20]. However, observing the calculation results in this paper, it is found that the classification scores of the three types of goods are significantly different. Most of them are distributed between 0.7-1. Hence, this paper uses the K-Means clustering algorithm in machine learning to take the sentiment scores of the reviews of the three types of goods as the data set and set up two clustering clusters to find the positive and negative classification points of the reviews. The final threshold is shown in Table 1, and the distribution of positive and negative reviews for the three categories of goods is shown in Table 2.

Table 1 The threshold for the sentiment classification of comments

Product category	Clothing	Laptops	Cosmetics
threshold value	0.70578	0.86637	0.77697

Table 2 The result of the distribution of positive and negative comments

Brand		Clothing	Laptops	Cosmetic
Number of comments		1462	1488	1808
Positive comments	Quantity	845	1129	1188
	Ratio	57.80%	75.87%	65.71%
Negative comments	Quantity	617	359	620
	Ratio	42.20%	24.13%	34.29%

3.4. TF-IDF Feature Word Extraction.

After calculating the sentiment score in this work, the retained dictionary is obtained and the weight of the input word sequence is calculated using the TF-IDF technique. The topic sequence terms are then derived from an analysis of the corpus's overall subject using the LDA topic model. The dictionary words are then contrasted with the LDA findings.

Filtering out meaningless terms as well as high-frequency and low-sensitivity terms is necessary because the preceding work will generate a lot of redundant terms. In addition, characteristics are added to the reserved word sequence to improve learning outcomes. The supplementary method is as follows: The topic word sequence of the documents produced by the corpus is determined using LDA, TF-IDF calculates the sequence, and a retained dictionary is formed by screening the low-frequency and high-discrimination terms. The terms in the dictionary are compared to the terms in the LDA results, and if they match, the reserved dictionary is expanded to include the sequence of subject words comprising the terms in the LDA results as input to the LDA model [21].

3.5. Cluster Analysis Based on LDA Topic Clustering Model

This study uses LDA thematic clustering on six groups of classified data (pos_Clothing, nes_Clothing, pos_Laptops, nes_Laptops, pos_Cosmetics, and nes_Cosmetics) to better understand the positive and negative aspects of customers.

3.5.1 Determination of the number of topics

Taking pos_Clothing as an example, before LDA topic clustering, you first need to determine the number of topics, and how to determine the number of topics in LDA has

yet to be recognized as a suitable method because different businesses have different requirements for generating topics. The most commonly used metrics in the industry include perplexity and coherence.

The molecular part of the Perplexity logarithmic function is the negative number that generates the likelihood estimation of the entire document set (indicating the generation ability of the parameters trained by the training set) because the probability value range is [0,1], according to the definition of the logarithmic function, the molecular value is a positive value and positively correlated with the text generation ability; The denominator is the number of words for the entire document set. Then, that is, the stronger the model generation ability, the smaller the Perplexity value, and the Perplexity calculation is as follows:

$$Perplexity(D_{test}) = exp\left\{\frac{-\sum_{d=1}^{M} log(p(w_d))}{\sum_{d=1}^{M} N_d}\right\} \tag{5}$$

Because the perplexity is not well applied in many scenarios, this paper combines the coherence of topic clustering to confirm the objectively optimal number of clustering topics. The coherence score was shown by David Mimno [22] to have a strong correlation with human judgment and can be used as an indicator to evaluate the topic model, and the larger its value, the better it is. The computation of coherence is divided into four main processes [23]: Segmentation, Probability Estimation, Confirmation Measure, and Aggregation, which finally can help us to better select the optimal number of topics, and the computation of Coherence is mainly done by The calculation of Coherence is mainly done by the gensim library in Python. The calculated perplexity curve and coherence curve are shown in Figure 5. (Take the negative comments on cosmetics as an example).

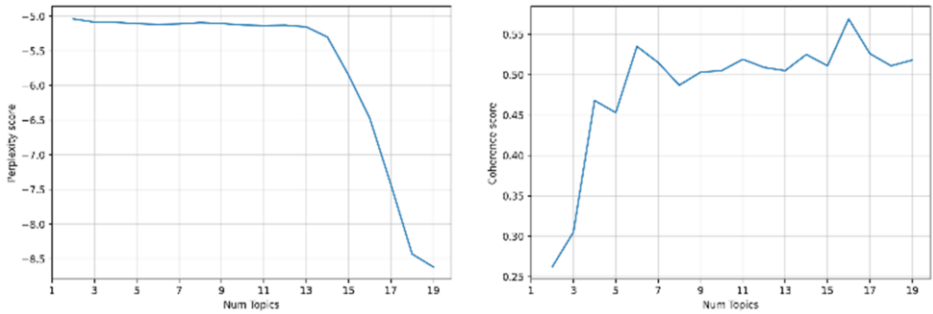

Figure 5 Perplexity curve and Coherence curve

In this paper, by clustering 1~20 topics in each data set, observing the image of their perplexity and coherence with the number of topics, the optimal number of topics is determined. However, because the number of optimal topics judged by perplexity and coherence is too objective, it may lead to overfitting and lead to unsatisfactory clustering effect, so this paper combines the visualization tools in the pyLDAvis library for a further callback. The optimal number of topics for each data set is finally determined, as shown in Table 3:

Table 3 Number of negative and positive topics by shopping category

	Clothing	Laptops	Cosmetics
Positive	5	4	6
Negative	4	5	5

3.5.2 Topic Clustering Results

After the above work, the positive and negative clustering results of three commodity categories can be obtained. Each topic can be subjectively named according to the frequency characteristics of the subject words. Each category's negative or positive decision-making factors can be explored through the distribution probability of the subject word. Since LDA topic clustering is a machine-learning model, it will inevitably overfit. Simple data preprocessing cannot dispose of 100% of useless information, so we combined it with the visualization module in the pyLDAvis library for further tuning.

Based on the above rules, the clustering results of the six types of datasets are shown in Table 4:

Table 4 Display of clustering results

Document	Topic	Words
Clothing(positive)	Style	Size Color Fit Design
	Texture	Effect Handle Details
	Brand	Black Style Serve
	Price	Quality Goods Activity Substantial Brick-and-mortar stores
	Workmanship	White Coherer Crozzling Burliness
Clothing(negative)	Style	Color Texture Fit Size
	Workmanship	Thrum Chromatism Yards cotton Variant
	Functionality	Breathability Sport Flaw Comfort level
	Serve	Customer service Attitude Consignment Price
Laptops (positive)	Game	Games Speed Keyboard System Experience Function
	Hardware	Appearance memory Solid disk Color gamut Processor
	Price	Price Shopping Activity Substantial
	Serve	After-sales System
Laptops (negative)	Game	Speed Starting-up Brothers Partner Refresh Keyboard
	Synthesis	After-sale Appearance Price Batteries
	Noise	Fan Vocality
	Function	Memory Solid Effect
	Screen	Light-leakage Flaw Video Camera Quality
Cosmetics (positive)	Entirety	Product Effect Color Taste
	Skin	Greasiness Skin Result
	Texture	Red Gift Hide
	Brand	Confide Wife Quality International Shoppe Experience
	Marketing	Giveaway Latex Oil-skin Suit
	Customer	Mudpack Dewy Vigor Object Cheap
Cosmetics (negative)	Product	Effect Activity Product Logistics Moisten Price
	Female	Color Lipstick Color-number Wife Gift Girlfriend
	Skin	Perceive Dry Mix
	Serve	Appraise service Speed Attitude Consignment Box
	Appearance	Complexion Appearance Collectivity

4. Results

4.1. Analysis of Negative and Positive Comprehensive Factors

To better understand the elements influencing consumers' positive and negative comments on clothing, computers, and cosmetics, respectively, a thematic clustering of the comments was done in the studies mentioned above. As a result, we shall compare and contrast the many emotional tendencies that fall under the same group.

4.1.1 Clothing

By looking at the theme categories for clothing items, we can observe that "style," "texture," "brand," and "workmanship" are the foundation for customers to give clothing products a positive rating, as well as the universal foundation for all clothing products. Further evidence that "style" and "service" are essential components of clothing products comes from the "brand" motif in positive comments. The brand may be the starting point for many consumers' purchase decisions before moving on to appearance and functionality. In other words, customers are enthusiastic about the company. For businesses or brands to dominate the market, increasing investment in brand marketing is a crucial step.

However, the interior features of "workmanship," "function," "service," and so on are where the majority of the bad parts of clothing products are located, customers using an online buying platform cannot learn anything about a product other than its appearance by physically handling it. Therefore, Manufacturers should implement more stringent controls over product quality and enhance product functionality based on the level of product quality. Enhance the product's ability to breathe and be comfortable; From the perspective of after-sales service, merchants should take the issue of returning after-sales service due to quality seriously. Additionally, developing a logistics platform with a multi-warehouse delivery mode may be a workable strategy to cut back on logistical expenses.

In sentiment analysis, negative comments account for the most significant proportion of 42.2% of clothing comments. Therefore, clothing products should focus more on users' pain points in negative comments to improve. The word clouds of positive and negative comments on clothing are shown in Figure 6 and Figure 7.

Figure 6 Positive Comments on Clothing

Figure 7 Passive Comments on Clothing

4.1.2 Laptops

According to the examination of comment data, young people—the majority of whom are either already enrolled in college or soon to be—are the main consumers of the laptop

product category on the JD platform. This group's primary traits include a solid education, good social skills, a passion for games, and a drive for productivity.

In the category of laptops, both the negative and positive clusters contain the theme of "game," which is mainly concentrated among the buyers of the game book. "Speed," "performance" and "keyboard" are essential bases for them to make good comments, while "speed" and "keyboard" are also the bases for making nasty comments. That is to say, for brands, no matter how much they strengthen the performance of hardware and the feel of the keyboard, they may be faced with a "less than flattering" situation because "performance" and "feel" will lead to the polarizing emotional tendency based on the subjective evaluation of different users. In this way, brands should not blindly enhance these two aspects. Instead, the allocation scheme that makes the marginal utility [24] as large as possible should be proposed based on the market preference and the allocation situation of other brands of the same type.

It is noteworthy that in this study, the negative comment data exhibits a singular clustering theme called "image," showing that although customers have high expectations for computing performance, they are also beginning to pay attention to image quality, including screen quality, camera quality, etc. The popularity of online workplaces and interviews has led to a gradual growth in consumer demand for this feature. The word clouds of positive and negative comments on laptops are shown in Figure 8 and Figure 9.

Figure 8 Positive Comments on Laptops

Figure 9 Passive Comments on Laptops

4.1.3 Cosmetics

The data study shows that women are more likely to be the target market for cosmetics. Additionally, under the impact of marketing, cosmetics have less price sensitivity and more important gift characteristics. Negative reviews tend to concentrate more on the product itself, including functionality, adaptability, and appearance aspects, whereas positive evaluations are dominated by marketing and the audience.

As a commodity with a high premium, brand image, and publicity are vital parts of cosmetics. Compared with the price, users pay more attention to whether the value of the brand meets the standard of gift giving and often do not consider its use value (as long as the color matching and quality control of the product are not too big problems). This standard is not unique, different brands will have different standards. Therefore will make the luxury category of brands often has its own research department, will based on consumer preferences build its own brand, their products to comprehensive packaging effect. It is clear from the unfavourable comments that the themes of "effect" and "function" are prevalent, thus it is crucial for online shoppers to have a positive experience with the cosmetics impact. The company should therefore focus more on the real impact of its products as well as the design and management of the beauty experience, in addition to improving its own publicity. It can create a more precise grouping of the target objects for the products and increase their suitability for various clients. The word

clouds of positive and negative comments on cosmetics are shown in Figure 10 and Figure 11.

Figure 10 Positive Comments on Cosmetics

Figure 11 Passive Comments on Cosmetics

4.2. Conclusion

This paper offers a strategy for platform operators and brand operators to analyze customer purchasing decisions by focusing on the comment data of three well-liked products on the JD platform. The novel part of this approach is quantifying the emotional tendency of comments using SnowNLP, a well-known Chinese corpus, categorizing the quantified comment data using enhanced K-Means machine learning, and combining TF-IDF feature word extraction with LDA topic clustering. This article offers recommendations for enhancing the platform and the brand by examining the precise causes of consumers' negative and positive attitudes. This paper contends that, in addition to the industries of clothing, laptops, and cosmetics, the aforementioned analysis method can be utilized in other commodity categories as long as the data are processed using the correct research techniques and the relevant vocabulary and context are understood.

Large corpora and big data have facilitated more efficient text analysis and given rise to numerous prominent corpora. The drawbacks of large corpora become clear when other scholars use the corpus for in-depth domain research: it can only perform sentiment analysis in mainstream contexts, while semantic analysis of subcultural and niche traffic circles leaves much to be desired and lacking, and researchers must have a strong grasp of particular semantic contexts. As a result, creating increasingly diversified context-specific text analysis corpora may represent a new study area.

References

[1] YOU Jun,ZHANG Xiaoyu,YANG Fengrui. (2019) Research on the influencing factors of the usefulness of online reviews: the regulatory effect based on commodity type[J]. *Soft Science*,2019,33(5): 140-144.)

[2] WEATHERS D, SHARMA S, WOOD S L (2007) Effects of online communication practices on consumer perceptions of performance uncertainty for search and experience goods［J］. *Journal of Retailing*,2007,83 (4): 393－401.

[3] Ajmal, A.; Aldabbas, H.; Amin, R.; Ibrar, S.; Alouffi, B.; Gheisari, M.J.C.I. (2022) Stress-Relieving Video Game and Its Effects: A POMS Case Study. *Comput. Intell. Neurosci.* 2022, 2022, 4239536. [CrossRef] [PubMed]

[4] LIU Y M, DU R. (2019) The effects of image-based online reviews on customers' perception across product type and gender［J］. *Journal of global information management*, 2019, 27 (3): 139－158.

[5] FINK L, ROSENFELD L, RAVID G. (2019) Longer online reviews are not necessarily better［J］. *International journal of information management*,2018, 39: 30－37.

[6] M. Hu and B. Liu. (2004) "Mining and summarizing customer reviews," in Proc. *the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2004, pp. 168–177.

[7] S. Mukherjee and P. Bhattacharyya. (2012) Feature Specific Sentiment. *Journal | [J]* CoRR. Volume abs/1209.2352, Issue.

[8] Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P (2011) Natural Language Processing (Almost) from Scratch.J. *Journal | [J]* CoRR. Volume abs/1103.0398, 2493–2537.
[9] David M. Blei, Andrew Y. Ng, Michael I. Jordan. (2003) Latent Dirichlet Allocation[J].*Journal of machine learning research*,2003,3(4/5). .source
[10] Basilio Marcio Pereira and Brum Gabrielle Souza and Pereira Valdecy. (2020) A model of policing strategy choice: The integration of the Latent Dirichlet Allocation (LDA) method with ELECTRE I[J]. *Journal of Modelling in Management*, 2020, ahead-of-print(ahead-of-print): 849-891.
[11] Zhou Mingyue and Yang Yu. (2021) Classificatory Analysis of Disputes on the Right of New Plant Varieties Based on LDA Model[J]. *Journal of Physics: Conference Series*, 2021, 1802(4): 042049-.
[12] Jiying Wu et al. (2020) A Competency Mining Method Based on Latent Dirichlet Allocation (LDA) Model[J]. *Journal of Physics: Conference Series*, 2020, 1682(1): 012059-.
[13] Wang Xuefeng et al. (2020) Evaluating the competitiveness of enterprise's technology based on LDA topic model[J]. *Technology Analysis & Strategic Management*, 2020, 32(2): 208-222.
[14] Feng Shaorong. (2020) A Microblog Unbalanced Data Evolution Analysis Method Based on LDA Model[J]. *Journal of Physics: Conference Series*, 2020, 1646(1): 012118-.
[15] Gang Han, Menggang Li, Yiduo Mei, Deming Li, Transportation Index Computation: A Development Theme Mining-Based Approach, *The Computer Journal*, Volume 64, Issue 3, March 2021, Pages 337–346, https://doi.org/10.1093/comjnl/bxaa102
[16] Muhammad Inaam ul Haq, Qianmu Li, Jun Hou, Analyzing the Research Trends of IoT Using Topic Modeling, *The Computer Journal*, Volume 65, Issue 10, October 2022, Pages 2589–2609, https://doi.org/10.1093/comjnl/bxab091
[17] Changxuan Wan et al. (2020) An association-constrained LDA model for joint extraction of product aspects and opinions[J]. *Information Sciences*, 2020, 519(C): 243-259.
[18] LIU Ce,LI Zhen,YAN Minghui.(2021) Research on text sentiment analysis for Dianping. com[J].*Modern Information Technology*,2021,5(19):37-39.
[19] Yang, Y., Xu, C., and Ren, G. Sentiment analysis of text using svm. In Wang, X., Wang, F., and Zhong, S. (eds.), Electrical, Information Engineering and Mechatronics 2011, London, pp. 1133–1139. Springer London.
[20] XIA Yuqin,SAN Xuewei. (2018) Simple text sentiment analysis based on Python [J]. *Yinshan Journal (Natural Science Edition)*, 2018, 32(4): 58-62
[21] FENG Yong, QU Bohao, XU Hongyan, et al. (2019) Chinese FastText Short Text Classification Method Based on TF-IDF and LDA[J].*Journal of Applied Sciences*,2019,37(03):378-388.)
[22] Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. *In Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).
[23] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. *In Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).
[24] LI Shuang. (2021) Research on the basic path of content e-commerce operation[J]*China Collective Economy*,2021,(30):108-109.