

Text Analysis of Danmaku in Orange Live Sales Based on LDA Model

Ziyi Zhang^{1,*}
Yitang Zeng²

*Corresponding author

¹ Wuhan Business University, Wuhan, China, 3203878529@qq.com

² Associate Professor, Wuhan Business University, Wuhan, China, 346316766@qq.com

ABSTRACT

Live e-commerce has revolutionized the marketing of agricultural products, but systematic analysis of unstructured live interactive comments by viewers remains underexplored. This study uses the Latent Dirichlet Allocation (LDA) model to analyze the latent topics of 770,000 comments in live broadcasts of orange sales at different merchant levels on the Douyin platform. By combining the axis-folding method with perplexity to determine the number of LDA topics, the analysis shows that: (1) product reviews emphasize product quality; (2) product prices highlight price sensitivity; (3) quasi-social interactions reflect emotional engagement; (4) user interests reflect purchase intentions; (5) e-commerce logistics focuses on delivery efficiency; and (6) the association between live broadcast promotions showing urgency prompts and conversion rates. The results empirically show that the success of agricultural product live broadcasts depends on product quality transparency, dynamic pricing, after-sales service, and emotional resonance between the host and the audience.

Keywords: LDA, topic modeling, live streaming, barrage, text analysis.

INTRODUCTION

Under the background of the in-depth implementation of the rural revitalization strategy, the "three rural" issues have always been the top priority of the work of the Party and the country. China's "Three Rural" Plan aims to promote the comprehensive upgrading of agriculture, the comprehensive progress of rural areas, and the comprehensive development of farmers. Live e-commerce, as an innovative model integrating the digital economy and rural industries, is becoming an important engine to help rural revitalization (Lv *et al.*, 2025). Oranges, as a typical fresh agricultural product, face pain points such as sales difficulties and information asymmetry. Live e-commerce, through real-time interaction and scenario-based display, has built a bridge for matching supply and demand between consumers and agricultural products such as oranges, and expanded the sales of agricultural products.

Currently, research in the field of agricultural product live broadcasts focuses on macro levels such as supply chain optimization and policy support, and in-depth mining of consumer interaction data is still insufficient. Live barrage, as an unstructured text of instant feedback from the audience, contains key information such as consumer preferences, emotional tendencies, and purchase decision logic. It is a "digital link" connecting the supply and demand sides of agricultural products (Xu & Ding, 2023). What are the focus of consumers on products such as oranges? The answers to these questions are of great practical significance for optimizing the operation strategy of agricultural product live broadcasts and promoting the implementation of the rural revitalization strategy.

Therefore, this study takes 770,000 bullet comments in the live broadcast rooms of orange sales at different levels on the Douyin platform as the analysis object, uses the latent Dirichlet allocation (LDA) topic model, combines the elbow method with perplexity calculation to determine the optimal number of topics, systematically mines the potential topic structure in the live broadcast bullet comments, and explores feasible paths for Internet technology to drive rural revitalization.

LITERATURE REVIEW AND THEORETICAL MODEL

Related Research

Latent Dirichlet Allocation (LDA), a classic probabilistic topic model, has demonstrated remarkable performance in uncovering the latent thematic structures within texts and has been extensively applied across diverse disciplines. In the healthcare domain, Danler *et al.* (Danler *et al.*, 2024) conducted a study titled "Visualizing Nursing Narratives: An Evaluation of the Latent Dirichlet Allocation Topic Model for Nursing Reports," validating the effectiveness of LDA in nursing text analysis. Meanwhile, Rajamaki *et al.* (Rajamaki *et al.*, 2024) utilized LDA in "Identifying Comorbidity Patterns in Alzheimer's Disease and Non-Alzheimer's Disease Patients," providing valuable data support for clinical diagnosis.

In the fields of social sciences and public opinion analysis, Oh and Kim (Oh & Kim, 2023) delved into the public perception of urology in South Korea through "Online News Data Analysis via Latent Dirichlet Allocation Topic Modeling." Zou *et al.* (Zou *et al.*, 2022) combined sentiment analysis with LDA in their research "Public Perception of Digital Fashion," offering insights into public attitudes towards digital fashion trends. In environmental science, Kim *et al.* (Kim *et al.*, 2021) employed LDA in "Determining Motivating Factors of Urban Forest Visitors using Latent Dirichlet Allocation Topic Modeling," effectively revealing the behavioral motivations of visitors.

Regarding methodological advancements, Chen *et al.* (Chen *et al.*, 2022) proposed the “Maximum Margin Deep Diversified Latent Dirichlet Allocation for Continual Learning” algorithm, enhancing the model’s performance in handling dynamic data. Hagg *et al.* (Hagg *et al.*, 2022) conducted a “Scoping Review of Analytic Practices of Latent Dirichlet Allocation in Psychological Science,” summarizing its application status in psychology. Hosoda *et al.* (Hosoda *et al.*, 2020) extended LDA’s application to microbiomics by “Revealing the Microbial Community Structure in the Human Gut Microbiome,” while Chen *et al.* (Chen *et al.*, 2017) adapted LDA for image processing with “Partial Membership Latent Dirichlet Allocation for Soft Image Segmentation.” Additionally, Hayashi (Hayashi, 2021) derived the “Exact Asymptotic Form of Bayesian Generalization Error in Latent Dirichlet Allocation,” providing a theoretical basis for model evaluation.

These studies collectively illustrate that LDA is not only effective for text data processing but also plays a pivotal role in interdisciplinary research. In this paper, taking the live-streaming sales bullet chat texts of oranges as the research object, we employed Python to crawl the bullet chat data from live-streaming rooms. The optimal number of LDA topics was determined using the elbow method in conjunction with perplexity. Subsequently, the LDA model was applied to organize and classify sentiment-laden texts, transforming product review texts into probabilistic topic distributions. The aim is to unveil the hidden thematic structures within the reviews, thereby providing actionable recommendations for consumer decision-making and business operations.

Principle of LDA Model

LDA is a probabilistic topic model based on an unsupervised machine learning algorithm. It assumes that a text is a mixture of multiple implicit topics, and each topic corresponds to a probability distribution of a set of words. Keywords are obtained through a three-layer Bayesian structure of “text-topic-word”. First, the topic distribution is sampled for each text, and then a topic is selected for each keyword based on the distribution. LDA is used as a conjugate prior of the multinomial distribution to characterize the prior information of topic distribution and keyword distribution. The LDA model uses the bag-of-words method, which treats the text as a word frequency vector, ignores the order of words, and converts the text into digital information. Each text corresponds to the probability distribution of a topic, and each topic corresponds to the probability distribution of keywords. Through Gibbs sampling, LDA can infer the implicit “text-topic” and “topic-word” distributions and the topic attribution of keywords from the text-word matrix.

The text layer contains a corpus of M texts, each text $d \in \{1, \dots, M\}$ is generated independently, the number of topics K is pre-set as a hyperparameter, and the parameter α controls the sparsity of the text-topic distribution. Each topic $k \in \{1, \dots, K\}$ in the topic layer is associated with a word multinomial distribution φ_k , and the parameter β controls the smoothness of the topic-word distribution. Each text d in the word layer contains N_d word positions, and each position $\{1, \dots, N_d\}$ corresponds to a latent topic variable $Z_{\{d, n\}}$. The model structure is shown in Figure 1.

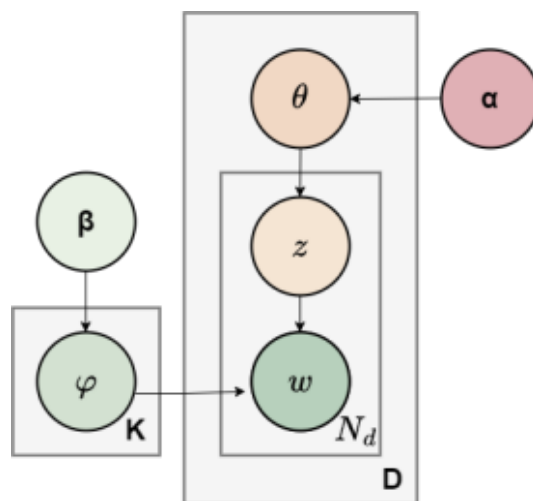


Figure 1: Schematic diagram of LDA model structure.

DATA COLLECTION AND PREPROCESSING

Data Collection Data Preprocessing

The data in this article was crawled from Douyin's orange sales live broadcast room. The bullet screen data was crawled for seven orange sales live broadcast rooms, namely the head, shoulder, middle, and tail, with a total of 770,000 live bullet screen data. The data was collected using a web crawler program written in Python. First, the Edge browser was driven based on the Selenium automated testing framework to obtain the unique identifier of the live broadcast room, room_id. Then, the live broadcast stream data channel was opened through WebDriver, and the Douyin bullet screen WebSocket protocol was reverse engineered for real-time bullet screen. Then, the websocket-client library was used to establish a long connection to listen to the bullet screen push, and the Protobuf technology was combined to process the bullet screen data packet. The structured fields

containing time, user nickname, and bullet screen text were extracted through a custom parser, and finally stored persistently in the MongoDB database in an incremental write mode, and a JSON log file with a time window marker was generated synchronously. The data set includes time, user nickname, and bullet screen text.

Data Preprocessing

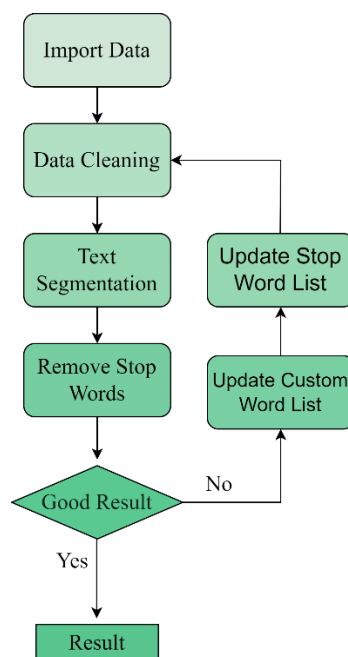


Figure 2: Data preprocessing flow chart.

The words in the text are closely connected. Compared with English, there are no separators. The text information is redundant and the word boundaries are complex. It is necessary to preprocess the text according to the content of the text and the research purpose to convert the unstructured data into structured data. The data preprocessing process is shown in Figure 2. First, the text is cleaned to delete interference symbols such as emoji expressions, special characters, and continuous punctuation marks. Delete meaningless numbers, call the re library, use regular expressions to replace punctuation marks and English letters with spaces, and then restore the homophonic characters or words to their original styles, such as "buoyancy" is restored to "welfare". Then call the jieba word segmentation library to perform part-of-speech tagging and keyword extraction on the barrage text, and import the self-built word library. According to the live broadcast context, retain network terms and emerging words to prevent words from being segmented and affecting the establishment of the model. After word segmentation, import the stop word list of Harbin Institute of Technology to remove stop words. Check the output results, manually correct the parts with inaccurate analysis or word segmentation errors, add the corrected words to the custom word library, and add irrelevant words to the stop word list. Repeat the above process until the effect is good. The results of text preprocessing are shown in Table 1.

Table 1: The reference styles in the text.

Original text	Processed text
I have a pair of black pants like this, they are super sticky!!	Black, pants, are, super, sticky
The strawberry greenhouse in front of my house also costs 40, 50 per pound, and there are several varieties, and the prices are different.	Strawberry, my house, greenhouse, 40, 50 pound, several varieties, prices
Ok, Tongma will help you apply for expedited delivery. Thank you for your support. Happy New Year!!	Tongma, apply for, expedited delivery. Thank you, support, Happy New Year
I wish you all a happy New Year. There is a lucky bag in the upper left corner.	happy New Year, lucky bag, in the upper left corne

ANALYSIS OF ORANGE LIVE COMMENTS BASED ON LDA TOPIC MODEL

Orange Review Analysis

The high-frequency words counted by word cloud are shown in Figure 3. "Delicious, sweet" reflects consumers' attention to taste. "Lucky bag, welfare" and other words show that consumers pay more attention to cost-effectiveness and discounts.

K-1 is an elbow point when the following conditions are met:

$$MI(K) < \delta \cdot \max_{t \leq K} MI(i) \quad (5)$$

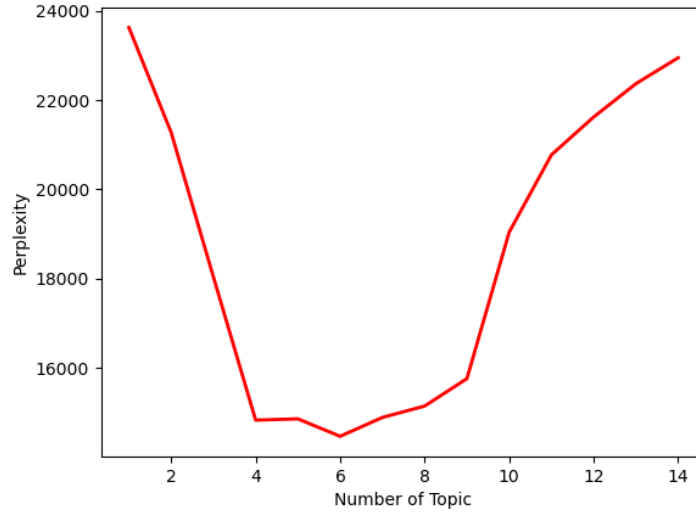


Figure 4: LDA topic number graph.

The perplexity function is called to calculate the perplexity under different numbers of topics, and the consistency of the number of topics is visualized through a line chart. When the perplexity is minimum within a certain range, the topics mined with a certain number of topics can better represent the topics of the data set. As can be seen from Figure 4, when the number of topics is 6, the perplexity is the lowest, and the model is best trained at this time.

Building an LDA Model

LDA assumes that each text is a mixture of multiple topics, and each topic is represented by a probability distribution of a set of words. For each text, a topic mixture ratio is extracted from the topic, and for each word in the text, a topic is extracted from the topic mixture ratio of the text, and a word is extracted from the word distribution corresponding to the topic. For each topic $k \in \{1, \dots, K\}$, $\phi_k \sim \text{Dirichlet}(\beta)$ is generated. For each text $d \in \{1, \dots, D\}$, $\theta_d \sim \text{Dirichlet}(\alpha)$ is generated. For each word $n \in \{1, \dots, N_d\}$ in text d , topics $z_{d,n} \sim \text{Multinomial}(\theta_d)$ and $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$ are generated.

The probability distribution of the LDA model is:

$$p(w, z, \theta, \phi | \alpha, \beta) = \prod_{k=1}^K p(\phi_k | \beta) \prod_{d=1}^D \left[p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{d,1} | \theta_d) p(w_{d,n} | z_{d,1}, \phi) \right] \quad (6)$$

w represents the set of all words, z represents the set of all topics, θ represents the topic distribution of all texts, and ϕ represents the word distribution of all topics.

In Gibbs sampling, the model is inferred by sampling the topic $z_{i,n}$ of each word, first fixing the topics of all other words, and then sampling the topic of the current word. The conditional probability is:

$$p(z_{d,n} = k | \mathbf{z}_{-(d,n)}, \mathbf{w}) \propto \frac{(n_{d,k}^{-(d,n)} + \alpha_k)}{\sum_{k'=1}^K (n_{d,k'}^{-(d,n)} + \alpha_{k'})} \cdot \frac{n_{k,w_{d,n}}^{-(d,n)} + \beta_{w_{d,n}}}{\sum_{v=1}^V (n_{k,v}^{-(d,n)} + \beta_v)} \quad (7)$$

$n_{d,k}^{-(d,k)}$ represents the number of words in text d assigned to topic k , $n_{k,v}^{-(d,v)}$ represents the number of times word v appears in topic k , α_k is the k th component of the text-topic Dirichlet distribution parameter α , and β_v is the v -th component of the topic-word Dirichlet distribution parameter β .

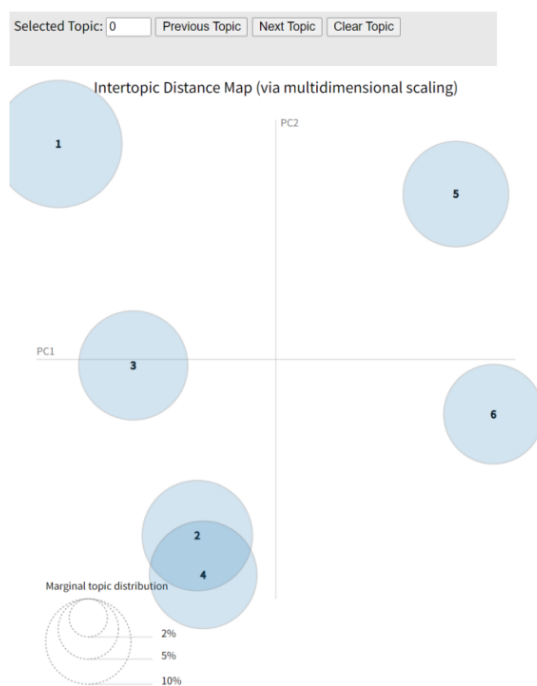


Figure 5: Theme and Characteristic Words.

The processed comment text is trained with the LDA model using the gensim library, and the number of topics K is set to 6. Finally, pyLDAvis is used for visualization analysis, and the distance of the circles represents the similarity between the topics. As shown in Figure 5, the distinction between topics 1, 3, 5, and 6 is obvious, and topics 2 and 4 overlap partially. Topic 2 is related to product prices, and topic 4 is related to user interests. Consumers' shopping interests are related to prices, and they usually make shopping decisions after asking about product prices. Therefore, it is reasonable that topics 2 and 4 overlap partially. The calculation results of the LDA model sort out and analyze the feature words corresponding to each topic, and use the top 15 feature words ranked by probability distribution to explain the topic, as shown in Table 2. This paper names the six topics of Orange Bullet Screen Comments as: product evaluation, product price, quasi-social interaction, user interest, e-commerce logistics, and anchor promotion.

Table 2: Topic Feature Words.

Topic	Keywords
Product reviews	Delicious, link, orange, video, size, what number, is there any, like, recommend, not sour, hot sale, cup, that kind, two, not bad.
Product prices	Welfare, delicious, this model, packaging, not expensive, situation, really, looking at, feeling, thing, express delivery, especially, live streaming room, taste, recommend
Quasi-social interactions	Snore, have a midnight snack, laugh to death, sleep, mukbang, grind teeth, rest assured, bench, cover one's face, be worth it, beautiful, at home, finish drinking, naughty kid, hairstyle
User interests	Paid, how large, ordered, recommend, repurchase, how large, purchased, acquire, how far, friend, hotpot, township, taste, discount, suitable
E-commerce logistics	Shipment, urgent, five to six days, address, delivery, please check, clothes, not good, tears, thank you, snicker, pick up, bag, freight insurance, is it good
Anchor promotion	Upper left corner, lucky bag, juicy, support, grab quickly, family members, direct supply from origin, follow, happiness, every day, pants, water cup, didn't watch, convenient

ANALYSIS OF ORANGE LIVE COMMENTS BASED ON LDA TOPIC MODEL

Product Reviews: Product Quality And Experience Drive Trust Building

Consumers' core focus on oranges is on the real experience and detailed information of the product. High-frequency words such as "delicious", "taste", and "not sour" indicate that consumers pay attention to the taste experience, indicating that taste is the primary factor affecting purchase decisions; words such as "size" and "number" reflect consumers' attention to fruit specifications and reflect their expectations for product standardization; the emergence of social words such as "recommended" and "hot selling" indicates that users rely on others' evaluation to assist in decision-making. In addition, words such as "link" and "video" indicate that consumers are accustomed to obtaining product information through intuitive content forms, while scenario-based words such as "cup" and "that" indicate that users are concerned about the adaptability of oranges in multiple scenarios such as family diet and beverage making. Therefore, merchants need to strengthen product quality control, establish a full-process traceability system from planting, warehousing and distribution, and display details such as orchard environment, picking process, packaging and delivery in real time through live broadcast rooms; encourage users to actively comment and check in, and use real experience word-of-mouth to improve trust; optimize product introduction pages, beautify interface design, and present product information in a visual way to reduce decision-making costs.

Product Price: Sensitivity to Price/Performance Ratio and the Desire for “Low Price But Not Low Quality” Coexist

Consumers are highly sensitive to prices and seek a balance between price and value, hoping for low prices without low quality. Words such as "welfare" and "exclusive price for live broadcasts" highlight the preference for purchase discounts, reflecting that the "instant discount" model of live broadcast e-commerce can stimulate consumers' desire to buy; positive comments such as "not expensive" and "cost-effective" appear together with words such as "packaging" and "express delivery", indicating that users expect low-priced goods to not sacrifice packaging texture and delivery efficiency; the appearance of subjective judgment words such as "look" and "feel" indicates that consumers will evaluate the rationality of prices through the content such as the live broadcast display screen and the appearance of the goods. If the packaging is simple and the fruit is damaged, it is inconsistent with the content displayed by the live broadcast, which is easy to cause negative comments. Merchants should adopt dynamic pricing strategies, set different preferential services in combination with inventory cycles and user portraits; optimize packaging design and product interface display to improve the unboxing experience within a controllable cost range; cooperate with logistics companies to provide "exclusive delivery services for live broadcasts", such as 48-hour expedited delivery, to strengthen the perception that low prices do not mean low quality.

Quasi-Social Interaction: Emotional Resonance is the Core Path to Building User Stickiness

Quasi-social interaction reflects the "social attributes" of live e-commerce, and consumers seek emotional resonance through the communication of anchors. The high frequency of life-related words such as "naughty child", "hairstyle", and "eating midnight snacks" shows that users want anchors to show their real personality, live broadcasts are close to the lives of the public, down-to-earth, rather than boring and stereotyped product salesmen; expression words such as "laughing to death" and "covering face" reflect the entertainment needs of interactive content, and relaxed and humorous words are more likely to resonate with consumers; words such as "rest assured" and "worth it" show that the anchor's response to user questions can significantly enhance consumer trust. The interaction between anchors and consumers not only prolongs the user's stay time, but also transfers the good impression of anchors to brands through emotional projection. Anchors need to strengthen their "personalized" positioning, design differentiated personalities, and cleverly integrate them into product sales in combination with life scenes; establish an interactive word library, and design professional and friendly replies to common orange usage and storage issues; carry out interesting activities to enhance the entertainment live broadcast experience.

User Interests: Deep Penetration of Social Recommendations and Decision-Making Links

User purchase decisions are highly dependent on social evidence and information verification. Behavior imitation words such as "taken" and "ordered" show that herd mentality is particularly prominent in live broadcast scenarios; words such as "how big", "how far", and "townships" reflect the cautious consideration of users in the sinking market on the delivery range and applicability; "repurchase" and "discount" reveal that the establishment of trust in the brand by old users requires continuous value feedback. The co-occurrence of words such as "friends" and "recommendations" shows that users are more inclined to reduce purchase risks through recommendations from friends and acquaintances. Merchants need to build a social fission mechanism, set up functions such as "share to receive discount coupons" and "invite friends to get gifts", and use consumer social circles to expand the product's dissemination radius; launch customized delivery services for different consumer levels to solve the "last mile" problem of delivery; establish a user tiered operation system, provide exclusive customer service, priority delivery and other privileges to high-repurchase users, and enhance brand loyalty.

E-commerce Logistics: Efficiency and After-Sales Service Determine the Minimum Experience

Logistics experience is a key factor affecting consumer satisfaction. Words such as "shipping", "expedited", and "five or six days" directly point to the timeliness of delivery. Consumers have a very low tolerance for "slow logistics", especially for fresh products such as oranges, which have a stronger demand for timeliness; words such as "freight insurance", "return", and "good or not" reflect consumers' emphasis on after-sales protection. Logistics delays or damaged goods can easily lead to negative comments from consumers and affect sales. Words such as "address" and "pick up" indicate consumers' demand for the convenience of receiving goods. Therefore, merchants prefer cold chain logistics for delivery, and establishing a standardized delivery process of "48-hour delivery + full temperature control" can better meet the needs of consumers, and relieve user anxiety.

by updating logistics nodes in real time; purchase freight insurance, simplify the return and exchange process, and frequently remind bad fruit compensation and other after-sales policies in the barrage.

Anchor Sales Promotion: Scenario-Based Marketing and Value Resonance Improve Conversion

The anchor's sales strategy needs to take into account both interest stimulation and emotional resonance. Words such as "lucky bag", "quick grab", and "upper left corner" reflect users' pursuit of instant rewards and merchants' strategies to attract consumers' attention. The limited time of only 50 orders can effectively create a sense of urgency and enhance consumers' desire to buy; words such as "direct supply from the origin" and "natural sweetness" meet consumers' demand for food safety and the trend of healthy living; affectionate names such as "family" and "support" strengthen identity recognition, while value words such as "happiness" and "original intention" enhance favorability through emotional resonance. Therefore, merchants can design a "limited order promotion" gameplay, combine countdown to enhance the sense of urgency of the opportunity; explore the cultural connotation of the product, and enhance the story of the product by telling the story of fruit farmers' planting, health concepts, etc.; optimize the rhythm of explanation, avoid intensive promotion information output, and adopt a three-stage structure of "knowledge popularization + experience sharing + limited time discount" to increase the content value density, so that live sales have warmth, technology, and knowledge.

CONCLUSION

The content of consumers' comments in the live broadcast room can reflect the degree of importance they attach to product characteristics and the direction of their attention. These characteristics often have an important impact on customers' purchasing decisions and can also help anchors improve their live broadcast style. Through the content of live broadcast barrage, we can get the focus of consumers' attention. This paper divides the topics and feature words into six categories: product evaluation, product price, quasi-social interaction, user interest, e-commerce logistics, and anchor promotion. In this study, consumers' attention to oranges is mainly concentrated in the following aspects: good taste, fresh ingredients, affordable price, rich nutrition, logistics and after-sales, and live broadcast fun. Therefore, anchors and merchants need to improve and improve the above aspects when selling agricultural products such as oranges, and formulate better operations and live broadcast sales styles. The text analysis method based on the LDA model can help us deeply understand consumer needs and improve the anchor's live broadcast method. The analysis results can provide merchants with effective improvement suggestions, optimize marketing strategies, improve sales performance, and increase user stickiness. The barrage data studied in this article only comes from Douyin and is only for the sales of oranges. It may not be applicable to other platforms. The research mainly focuses on thematic classification and does not further explore the sentiment of keywords. In the future, it can be considered to include multi-platform and multi-category agricultural product data to analyze the commonalities and differences of consumer barrage data across scenarios.

ACKNOWLEDGMENT

This study is supported by the Guiding Program of Philosophy and Social Sciences for Higher Education Institutions in Hubei Province (prefunded program of Provincial Social Science Foundation) (Grant No.21G168), and the Team for High-quality Development of Cross-border E-commerce in Wuhan Business University.

REFERENCES

- Chen, C., Zare, A., Trinh, H. N., Omotara, G. O., Cobb, J. T., & Lagaunne, T. A. (2017, 1201). Partial Membership Latent Dirichlet Allocation for Soft Image Segmentation [Article]. *IEEE Trans Image Process*, 26(12), 5590-5602. <https://doi.org/10.1109/tip.2017.2736419>
- Chen, W., Chen, B., Liu, Y., Cao, X., Zhao, Q., Zhang, H., & Tian, L. (2022, 0715). Max-Margin Deep Diverse Latent Dirichlet Allocation With Continual Learning [Article]. *IEEE Trans Cybern*, 52(7), 5639-5653. <https://doi.org/10.1109/tcyb.2020.3044915>
- Danler, M., Hackl, W. O., Neururer, S. B., Huber, L., & Pfeifer, B. (2024, 0822). Visualizing Nursing Narratives: An Evaluation of Latent Dirichlet Allocation Topic Modeling for Care Reports [Article]. *Stud Health Technol Inform*, 316, 1709-1713. <https://doi.org/10.3233/shti240752>
- Hagg, L. J., Merkouris, S. S., O'Dea, G. A., Francis, L. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Westrupp, E. M., Macdonald, J. A., & Youssef, G. J. (2022, 1108). Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review [Article]. *J Med Internet Res*, 24(11), e33166. <https://doi.org/10.2196/33166>
- Hayashi, N. (2021, 0501). The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation [Article]. *Neural Netw*, 137, 127-137. <https://doi.org/10.1016/j.neunet.2021.01.024>
- Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M., & Hamada, M. (2020, 0623). Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation [Article]. *Microbiome*, 8(1), 95. <https://doi.org/10.1186/s40168-020-00864-3>
- Kim, D. S., Lee, B. C., & Park, K. H. (2021, 0913). Determination of Motivating Factors of Urban Forest Visitors through Latent Dirichlet Allocation Topic Modeling [Article]. *Int J Environ Res Public Health*, 18(18). <https://doi.org/10.3390/ijerph18189649>
- Lv, J., Guo, X., & Jiang, H. (2025). Rural E-Commerce and Income Inequality: Evidence from China. *Sustainability*, 17(10), 4720.
- Oh, Y. W., & Kim, J. (2023, 1115). Insights Into Korean Public Perspectives on Urology: Online News Data Analytics Through Latent Dirichlet Allocation Topic Modeling [Article]. *Int Neurourol J*, 27(Suppl 2), S91-98. <https://doi.org/10.5213/inj.2346288.144>

- Rajamaki, B., Braithwaite, B., Hartikainen, S., & Tolppanen, A. M. (2024, 0615). Identifying Comorbidity Patterns in People with and without Alzheimer's Disease Using Latent Dirichlet Allocation [Article]. *J Alzheimers Dis*, *101*(4), 1393-1403. <https://doi.org/10.3233/jad-240490>
- Xu, R., & Ding, B. (2023). An empirical study on purchasing intention of agricultural product e-commerce live broadcast. SHS Web of Conferences,
- Zou, Y., Luh, D. B., & Lu, S. (2022, 0615). Public perceptions of digital fashion: An analysis of sentiment and Latent Dirichlet Allocation topic modeling [Article]. *Front Psychol*, *13*, 986838. <https://doi.org/10.3389/fpsyg.2022.986838>