Consumer Online Review Content Mining and Emotion Analysis Under the Background of Live Streaming E-Commerce

Fang Zhang [1] Fengxiao Li [2*]

（School of Marketing Management, Liaoning Technology University, Huludao, Liaoning, 125105, China） [1,2]

zhangfang@Intu.edu.cn[1] 15904294871@163.com [2*]）

Abstract. Online reviews of products in the mode of live e-commerce contain important factors that affect consumers' purchasing decisions. The content of agricultural product reviews on the Tiktok platform was collected by Python web crawler, and 23557 agricultural product reviews were analyzed by word frequency analysis, network co-occurrence of high-frequency word meanings, LDA theme model, SnowNLP sentiment analysis and other methods. The study found that product quality, express packaging, anchor recommendation and logistics speed are the key factors affecting consumers' online purchase decisions under the live streaming e-commerce model. On this basis, strengthening product quality, improving express packaging, improving the professionalism of anchors, and ensuring logistics speed are important measures to improve consumers' desire to buy, and are of great significance to promote the sustainable development of the agricultural live e-commerce industry.

Keywords： electricity supplier logistics; LDA topic model; live-streaming e-commerce; agricultural products; SnowNLP sentiment analysis

1   INTRODUCTION

According to the 51st Statistical Report on China's Internet Development released by China Internet Network Information Center (CNNIC), as of December 2022, the number of Internet users in China has reached 1.067 billion, and the Internet penetration rate has reached 75.6%[1]. With the diversified development of online shopping forms, live streaming e-commerce platforms can present the planting, processing and other links of agricultural products to consumers through live streaming, and establish a trust relationship between products and consumers. Consumers' demand for online purchasing of agricultural products is constantly increasing, which brings opportunities for the development of e-commerce live streaming industry of agricultural products. As a new e-commerce poverty alleviation model, live broadcast of agricultural products has achieved significant economic and social benefits in terms of increasing farmers'

income and helping poor households get out of poverty [2]. Tiktok is one of the leading platforms in China's e-commerce livestream industry. In order to enable the operators of agricultural e-commerce livestream to better understand the factors affecting consumers' purchase, this paper uses Python web crawler technology to obtain the online comment data of the top three products sold in the commodity window of the product account of Tiktok's agricultural livestream. Word frequency statistics, LDA theme model, SnowNLP sentiment analysis and other methods were used to extract the key factors affecting consumer purchase, and provide reference suggestions for promoting the sustainable development of agricultural products e-commerce and logistics.

2   LITERATURE REVIEW

Xuchao Song adopted the empirical research method, selected e-commerce live broadcasting as the research object, took fan economy as the intermediary variable, revealed the influence mechanism of e-commerce live broadcasting on agricultural product marketing, and reached the conclusion that e-commerce live broadcasting and fan economy have a positive impact on agricultural product marketing [3]. Based on the characteristics of live streaming, Hongpeng Guo et al. analyzed the impact of perceived risk and value co-creation on consumers' purchase intention [4]. Through questionnaire survey, Yongjia Chang revealed the conclusion that live marketing and perceived value of agricultural products have a significant positive impact on consumers' purchase intention [5]. Based on the stimulus-body-response model, Dawei Liu et al. found that the influence of anchors in the process of live broadcasting can improve consumers' purchase intention [6].

Fengcai Ma et al. measured the satisfaction of fresh products sold on e-commerce platforms and put forward relevant suggestions by means of word frequency statistics and construction of feature word lists [7]. Hongbo Tu et al. found that perceived risk had a significant negative impact on consumers' willingness to buy geographical indication agricultural products online. In order to attract consumers to buy, merchants should pay attention to the content and number of positive reviews [8].

Most of the above studies are based on questionnaires or user comments of Jingdong Mall as data support. The difference of this study is that it takes commodity comments of Tiktok platform, which has a large traffic volume in China, as data source, and uses LDA subject modeling, emotion analysis and other methods to explore the main characteristics of consumer evaluation content and user emotion under the live streaming e-commerce model. Suggestions for agricultural products live e-commerce operators.

3   RESEARCH METHODS AND DATA COLLECTION

3.1   Research Framework

The research framework of this paper is shown in Figure 1 below. A web crawler written in Python is used to collect the top three product reviews of the merchandise window sales of the agricultural products live account on the Tiktok platform as the initial

analysis text, and the data are de-processed to eliminate invalid data and manually delete the special expressions and symbols in the data. The "jieba" word segmentation package in Python is used for word segmentation processing, filtering and stopping words, etc. The content characteristics of the data are analyzed by word frequency analysis, network co-occurrence analysis of high-frequency word meaning and LDA topic model analysis, and the comments of agricultural products consumers under the background of live streaming e-commerce are mined. The trained SnowNLP library was used to conduct emotion analysis on the whole text to understand consumers' emotional tendency of evaluating agricultural products.
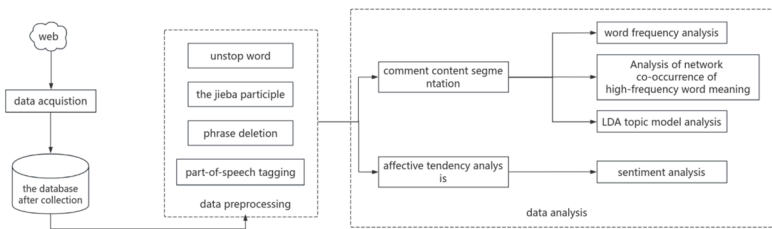

Fig. 1. Research framework

3.2   LDA Topic Clustering Model

LDA (Potential Dirichlet Assignment) model is a generative Bayesian probability model proposed by Blei et al., which is widely used in the field of text mining [9]. It provides a way to analyze and mine hidden structural information in text without having to annotate the data beforehand. The basic idea of LDA model is to regard text data as composed of three levels: Word, Topic and corpus. Each document represents a probability distribution for a set of topics, and each topic represents a probability distribution for a set of words. By fitting the LDA model, we can get the core keywords of each topic and the probability of occurrence.

3.3   Data Collection

By writing a web crawler in Python, the review content of the top three products sold in the shop window of the agricultural products live account of the Douyin platform was collected, and a total of 24,031 comments were collected on August 9, 2023. The collected data includes comment user name, comment content, comment time, comment likes, comment score and other information. After weeding out duplicate and invalid data (such as "default praise") and manually removing special emojis and symbols, we ended up with 23,557 valid comments.

4   RESEARCH PROCESS AND DATA ANALYSIS

4.1   Analysis of Consumer Review Content

High-frequency Word Analysis. Synonym substitution is performed on the initial text content of consumer reviews to remove ambiguity and ensure the accuracy of word frequency statistics. According to the evaluation content of consumers, a specific custom dictionary is built, and mainstream stop word lists such as "Hit stop word Database" are used to eliminate words that frequently appear in the evaluation but have no practical significance, and words with little reference value are manually added to the stop word list to ensure that the final data has a high reference value. The evaluation content of consumers is processed by using the "jieba" word segmentation package in Python.

Table 1. Top10 high-frequency words

category	High frequency noun		High frequency verbs		High frequency adjective words	
ranking	keyword	word frequency	keyword	word frequency	keyword	word frequency
one	speed	6673	buy-back	6509	good	2436
two	logistics	5144	deliciousness	6473	integrity	2278
three	drikold	4437	pack	4584	compaction	544
four	size	4005	shipments	3743	freshness	472
five	children	3763	receive	3332	neat	448
six	ice block	3484	like	2705	indeed	414
seven	taste	3428	not melt	1951	briefness	315
eight	quality	3292	description	1617	delicious	280
nine	flavour	2538	correspond	1502	excellent	248
ten	direct broadcasting room	2131	deserve	1198	fresh and sweet	237

The higher the frequency of the keyword, the more consumers pay attention to the keyword. According to the word cloud map and the high-frequency word list, it can be seen that the agricultural consumers of live streaming e-commerce are most concerned about "speed" after purchasing products. Timely delivery and fast logistics can ensure the quality and taste of agricultural products, meet the immediate needs of consumers, and enhance consumers' shopping experience and satisfaction. In the list of high-frequency words in nouns, agricultural consumers focus on logistics, product quality, cooling materials, etc., many users evaluate it as "logistics speed is fast" and "quality is good", but there are also some feedback comments on issues such as "ice pack leakage". In the list of high frequency verbs, repo becomes the key word. When consumers buy agricultural products through the recommendation of anchors, friends planting grass, etc., if the agricultural products meet the description of the anchor or the psychological expectations of consumers, they may have the idea of buying back. In the list of adjective high frequency words, by using the database collected by combining high frequency words, it is found that there are many descriptions of "good taste" and "fresh and tender meat" in consumers' product reviews. In addition, in the list of high-frequency adjective words, there are no words expressing negative emotions, and the overall comments are positive.

Co-occurrence Analysis Based on Semantic Network. ROST CM 6 software developed by Wuhan University was used to divide the cleaned data into words, and the top 30 high-frequency words in commodity reviews were counted. By sorting out the frequency of pairwise co-occurrence of these 30 high-frequency words, the high-frequency word co-occurrence matrix as shown in Table 2 is constructed.

Table 2. High-frequency word co-occurrence matrix (part)

	speed	delicious-ness	buy-back	drikold	pack	logis-tics	size	ship-ments	. . .
speed		503	455	915	1596	2137	525	2213	. . .
delicious-ness	503		2015	859	798	758	468	683	. . .
buy-back	455	2015		996	808	841	928	717	. . .
drikold	915	859	996		1891	1434	1082	1147	. . .
pack	1596	798	808	1891		1993	642	1765	. . .
logistics	2137	758	841	1434	1993		809	1727	. . .
size	525	468	928	1082	642	809		654	. . .
shipments	2213	683	717	1147	1765	1727	654		. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

By analyzing the co-occurrence matrix, it can be found that consumer reviews form a review network with "speed" as the core. Among them, packaging, logistics, delivery and other words have a higher number of co-occurrences with other high-frequency words, showing the close correlation between these words and other high-frequency words. From the whole co-occurrence matrix, the two closely related parts are summarized as product packaging and logistics speed. In terms of express packaging, consumers pay special attention to terms such as "packaging", "dry ice" and "ice" when purchasing agricultural products through live streaming, highlighting the importance consumers attach to express packaging, which shows that consumers hope that agricultural products can be packaged and preserved in a proper way to ensure product quality and taste; In terms of logistics speed, terms such as "logistics" and "delivery" highlight consumers' pursuit of logistics speed and delivery speed, and consumers hope to receive the purchased agricultural products as soon as possible to ensure the freshness of products.

Feature Analysis Based on LDA Topic Model. Compared with the degree of confusion, topic consistency can more accurately evaluate the effect of topic extraction in LDA model. When the topic consistency score reaches the maximum value, it means that the result of topic extraction is the most reasonable [10]. Therefore, this paper uses the topic consistency score to determine the optimal number of topics. As can be seen from Figure 2, when the number of topics is 4, the consistency score is the highest, and with the increase of the number of topics, the consistency score gradually becomes stable, so the optimal number of topics in this paper is 4. The top 10 feature words in each topic are selected as examples and presented in Table 3.

In order to determine the topic name of the LDA model, two groups of researchers were invited to analyze the topic name generated by the LDA model. By integrating the summarized results of the two groups of researchers and combing the previous studies on agricultural products under the live streaming e-commerce model [11-13], the first topic was product quality. The second theme is express packaging; Theme three recommendation of the main broadcast; The fourth theme is logistics speed.
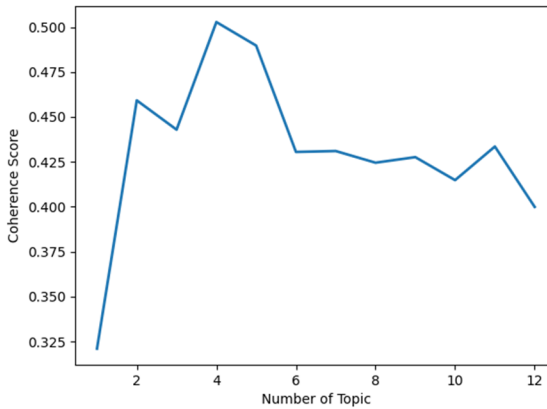

Fig. 2. Topic consistency curve

Table 3. LDA subject classification and subject word probability

product quality		express packaging		anchor recommendation		the speed of logistics	
deliciousness	0.096	drikold	0.099	size	0.06	logistics	0.092
buy-back	0.06	ice block	0.069	buy-back	0.037	fast speed	0.075
children	0.053	express	0.038	real	0.036	shipments	0.068
like	0.04	not melt	0.036	quality	0.032	speed	0.065
taste	0.034	ice pack	0.035	deserve	0.021	pack	0.061
flavour	0.029	place an order	0.029	good	0.021	cost performance	0.049
meat quality	0.022	pack	0.025	recommend	0.018	true to form	0.048
real	0.018	yesterday	0.012	anchor	0.018	description	0.034
especially	0.016	logistics	0.012	direct broad-casting room	0.016	customer service	0.033
quality	0.014	a little	0.011	stuff	0.016	correspond	0.032

The analysis results of each potential topic in the LDA topic model are as follows:

First, product quality mainly includes taste, taste, meat quality and so on. According to consumer comments, agricultural products purchased under the live e-commerce model are not only good quality and cheap, but also loved by their families.

Second, for express packaging, consumers are concerned about whether the packaging is intact. Especially for products that are easily corrupted or need to be kept fresh, consumers are more concerned about whether the express is shipped by cold chain, and whether there are cooling materials such as dry ice and ice cubes in the package. Express packaging also includes product packaging [14]. For product packaging, consumers often mention in their comments whether the purchased products have independent packaging, whether the packaging is exquisite and whether the product quality can be effectively protected.

Third, in the live e-commerce environment, the degree of recommendation of anchors for agricultural products is one of the important factors for consumers to purchase for the first time. Consumers often mention content such as "believe the recommendation of xx anchor" in the comments, which indicates the dependence of consumers on the recommendation of anchor when purchasing agricultural products. In addition, the performance of anchors in the live broadcast process has an important impact on consumers' purchase decisions and subsequent behaviors.

Fourth, the speed of logistics is the key factor mentioned repeatedly by consumers in the comments. In the process of the distribution of agricultural products, consumers will mention and praise the door-to-door express delivery and couriers with good service attitude in the product reviews. In live streaming e-commerce, improving logistics speed and providing door-to-door delivery services is one of the important factors to attract consumers and increase consumers' willingness to buy.

4.2   Sentiment Analysis of Consumer Reviews

Emotional analysis of consumer review text can understand consumers' emotional attitude towards agricultural products purchase under the live e-commerce model. Use the SnowNLP library in Python for sentiment analysis. The sentiment analysis model of SnowNLP is trained based on online comments of e-commerce, and the output range is between [0,1]. The distribution of consumers' emotional propensity to comment is shown in Figure 3. It can be seen from the figure that consumers generally make more positive comments on the evaluation distribution of agricultural products purchased live.
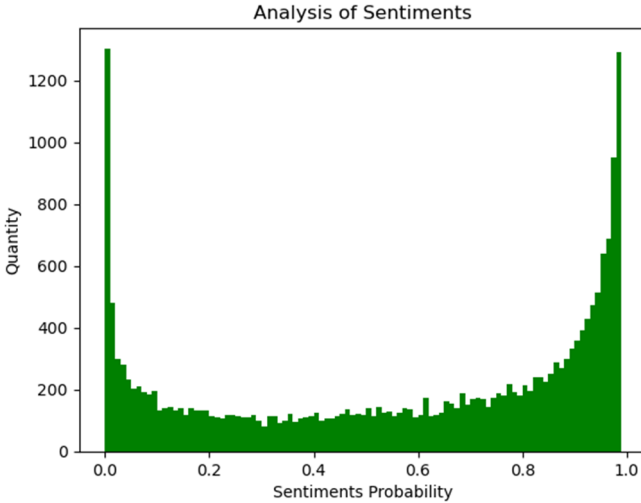
Fig. 3. Emotional distribution

5   CONCLUSIONS AND SUGGESTIONS

This paper selects commodity review data of livestream accounts of agricultural products with goods on Douyin platform, and uses word frequency analysis, network co-occurrence of high-frequency word meanings, LDA theme model, and SnowNLP emotion analysis methods to dig out the characteristics of consumer review content and user emotion under the background of livestream e-commerce. The main conclusions are as follows:

(1) The factors affecting consumer purchase in the live streaming e-commerce model mainly include product quality, express package, anchor recommendation, and logistics speed.

(2) According to the SnowNLP emotional distribution map, some negative comments exist, indicating that relevant services need to be improved under the live broadcast mode of agricultural products e-commerce. The content of comments on agricultural products is the focus of consumers. Therefore, suggestions for improving the four themes contained in comments on agricultural products extracted by LDA thematic model are put forward:

① In terms of product quality, before delivery, agricultural products e-commerce operators hire experienced professionals to ensure product quality from the source of the product, and implement strict quality management processes.

② Different types of agricultural products have different packaging requirements and transportation needs, businesses need to customize appropriate packaging for different types of products, enhance the image of agricultural products, increase consumer satisfaction with agricultural products, reduce the probability of consumers to make bad reviews.

③ Anchors play an important role in the live e-commerce model and are a key factor in promoting consumer purchases. During the livestreaming process, anchors can tell the stories behind agricultural products and brand culture, connect agricultural products with consumers' emotions, resonate and stimulate the desire to buy. In addition, when selecting products, anchors can select some products from poor areas with high quality and easy to sell, helping farmers in poor areas to achieve income increase.

④ In terms of logistics, the delivery time and expected delivery time of agricultural products are the concerns of consumers, and they are also often mentioned by consumers in product reviews and broadcast rooms, so agricultural live e-commerce operators take appropriate logistics measures by choosing reliable logistics partners to ensure logistics speed.

The limitation of this paper is that it only selects the data of agricultural products livestreaming accounts on the Douyin platform without considering the influence of time factor. In the future, the scope of research can be further expanded and time factor can be included in the research content.

REFERENCE

1. China Internet Network Information Center releases the 51st Statistical Report on Internet Development in China [J]. Journal of the National Library, 2019,32(02):39.
2. Mengying Zan, Zhengbing Wang. Live broadcast of agricultural products e-commerce: a new model of poverty alleviation through e-commerce [J]. Issues in Agricultural Economy,2020(11):77-86.
3. Xuchao Song. Analysis of the influence mechanism of e-commerce live broadcast on agricultural product marketing--Based on the mediating role of fan economy [J]. Business Economics Research,2023(07):80-83.
4. Hongpeng Guo, Xiangnan Sun, Chulin Pan, et al The Sustainability of Fresh Agricultural Produce Live Broadcast Development: Influence on Consumer Purchase Intentions Based on Live Broadcast Characteristics[J]. Sustainability, 2022, 14 (12):7159-7178.
5. Yongjia Chang. Study on the impact of livestream of agricultural products on consumers' purchase intention -- Based on the transmission effect of perceived value and perceived risk [J]. Valence Theory and Practice,2022(07):130-133.
6. Dawei Liu, Jinyang Yu. Impact of perceived diagnosticity on live streams and consumer purchase intention: streamer type, product type, and brand awareness as moderators.[J]. Information technology & management,2022: 11-14.
7. Fengcai Ma, Chunyue Li. Measurement of consumers' satisfaction with fresh products sold on e-commerce platform: An analysis based on Jingdong Fresh Online review [J]. Price Theory and Practice,2020(05):117-120.
8. Hongbo Tu, Qingjian Zhu. The influence and mechanism of external cues on the willingness of geographical indication agricultural products to shop online [J]. China Circulation Economy,2020,34(08):37-47.
9. Blei D M,Ng A Y,Jordan M. Latent dirichlet allocation[J].Journal of Machine Learning Research,2003,3:993-1002.
10. Jianhong Ma, Chenxi Wang, Yan Lin et al. Synthesis and analysis of specialty technology subject based on product life cycle [J]. Chinese Journal of Information Technology, 2012,41(07):684-691.

11. Huamin Ma. Research on the upgrading path of marketing model of agricultural products promoted by livestreaming delivery [J]. Research of Commercial Economics,2021(15):81-84.
12. Yue Xu, Fuyuan Zheng, Weiping Chen. Influence of technology availability and anchor characteristics on consumers' purchase intention of agricultural products [J]. Rural Economy,2021(11):104-113.
13. Zhao Li, Jianming Sun, Haifeng Zhao et al. Packaging design of Yongquan Satsuma Express [J]. Packaging Engineering,2018,39(21):121-125.
14. Lei Ma. Integrated packaging design of online shopping products based on commodity classification attributes [J]. Packaging Engineering,2017,38(14):32-36.