Analysis of Co-branded Food Product Reviews Based on BERTopic and SnowNLP

Dai Wei

School of Computer and Information Engineering, Hubei University, Wuhan, Hubei, 430062, China
wdai@stu.hubu.edu.cn

Abstract: *In 2023, the scale of the co-branding economy in China surpassed 100 billion yuan, demonstrating a rapid growth trend, with more than half of the popular co-branded products belonging to the catering industry. However, alongside the apparent rapid development of the co-branding economy, many negative consumer reviews have also gradually surfaced, raising the issue of how to sustain high-quality development in the co-branding economy as a critical concern today. This paper addresses this issue by using Python's weibopy library to crawl comments from Sina Weibo, selecting comments on four popular co-branded products, including Luckin Coffee × Jackson Yee. To delve deeply into the themes within the comments, the BERTopic topic model is used to extract seven representative co-branding themes, revealing that packaging and celebrities are of significant concern to consumers. Finally, SnowNLP sentiment analysis is applied to understand consumer sentiment in four of these themes, leading to suggestions that co-branded products should consider addressing negative feedback regarding taste.*

Keywords: *Web scraping; BERTopic; SnowNLP; Sentiment analysis; Catering co-branding*

1. Introduction

The co-branding economic model refers to a business strategy where brands collaborate with other brands, designers, or artists to jointly develop products, enhance brand recognition, and achieve mutual benefits through co-branding cooperation[1]. In recent years, the catering industry has seen a surge of popular cross-industry co-branding cases, becoming a hot spot in brand marketing. Starting from the first half of 2023, several high-profile co-branding cases have emerged, including collaborations between Fendi and Heytea, and Moutai and Luckin Coffee. The continuous emergence of co-branding cases has started to make the public feel indifferent, and the prolonged development of a single form of co-branding may lead to consumer aesthetic fatigue. Only by elevating co-branding marketing to a higher level of strategic thinking to support the long-term development of brands can the power of the co-branding economy continue to be exerted[2].

To gain an in-depth understanding of the current state of the co-branding economy in the catering industry, this paper employs web scraping and natural language processing techniques. Using the weibopy library, we obtained textual comments from Sina Weibo, preprocessed the textual data. Subsequently, the BERTopic model was used to select representative themes, and sentiment analysis was conducted for each theme. Through comprehensive analysis, this paper proposes more substantial development suggestions for the co-branding economy in the catering industry to promote its positive development.

2. Data Acquisition and Preprocessing

2.1. Data Sources

To more comprehensively study the response to different types of co-branded catering economies on social media platforms, this paper selects Sina Weibo as the platform for data crawling. Weibo is one of China's popular social media platforms, boasting a vast user base. Based on user comments and interactions on Weibo, this paper uses text mining methods to analyze the overall situation of the co-branded catering economy[3].

For data acquisition, this paper calls relevant interfaces through the Weibo API and uses Python's weibopy library to crawl online comments. Using the popularity of the co-branded catering economy on Weibo as a reference, this paper selects four highly popular and successful co-branding cases for text comment crawling, retrieving 500 comments for each case. The four successful co-branding cases are Luckin Coffee × Jackson Yee, Heytea × Fendi, Luckin Coffee × Moutai, and Pizza Hut × Genshin Impact.

2.2. Data Preprocessing

Before conducting text data mining, data preprocessing is a crucial step to ensure the accuracy and interpretability of subsequent results. Online comments are diverse in form and may contain unstructured elements such as emojis and internet slang, which reduce data standardization and may lead to distorted results if analysis methods are applied directly. Therefore, data preprocessing using Python libraries such as Pandas and jieba is conducted, including tokenization, removal of stop words, and elimination of emojis[4]. These operations effectively clean noise and extract key information, providing high-quality input for subsequent text mining and topic modeling.The text data preprocessing workflow is illustrated in Figure 1.

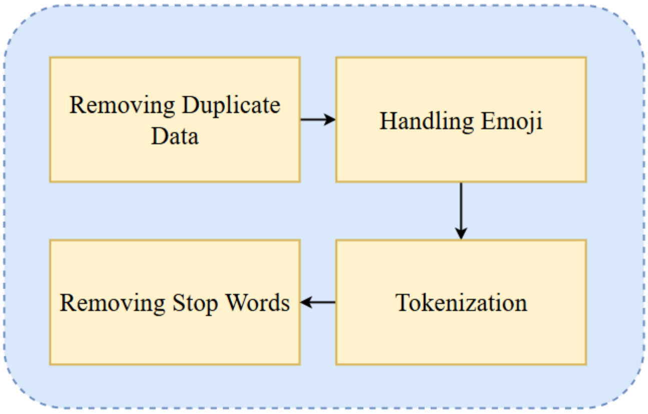

Figure 1: Text Mining Preprocessing Flowchart

The specific process of data cleaning and preprocessing is as follows:

2.2.1. Removing Duplicate Data

To ensure the uniqueness and accuracy of the data, this study uses the "drop_duplicates()" method from the Pandas library to remove duplicate comments. The presence of duplicate data can lead to biases in statistical results and affect the reliability of the analysis. By removing duplicate comments, we can obtain a more authentic and reliable dataset, ensuring that subsequent topic modeling and sentiment analysis are not influenced by duplicate data.

2.2.2. Handling Emoji

Comment data often contains various emojis, which, while enriching expression, can pose challenges in natural language processing. To address this issue, this study employs regular expressions to identify and handle emojis in the comments. This approach allows for the effective extraction or removal of emojis, making the text processing more standardized.

2.2.3. Tokenization

In Chinese natural language processing, tokenization is a fundamental and critical step. This study uses the well-known Chinese tokenization tool jieba to split the text into individual words. The jieba tokenization tool is implemented based on a Trie tree structure and prefix dictionary, and combines word frequency statistics with rules for Chinese word formation. It provides efficient and accurate Chinese tokenization. The results of tokenization serve as the foundation for subsequent text analysis and feature extraction, directly impacting the performance of models and the accuracy of analysis results.

2.2.4. Removing Stop Words

Stop words are those that occur frequently in a given text but carry little meaningful information. To enhance the effectiveness of text analysis, this study defines a set of common stop words and stores them in a text file. By removing these stop words during data preprocessing, key information in the text is preserved, and noise is reduced. This process helps improve the performance of tasks such as text classification, clustering, and sentiment analysis.

3. Topic Selection Based on BERTopic Model

3.1. Construction of the BERTopic Model

BERTopic is a topic modeling technique based on the BERT model and is an unsupervised learning method that automatically discovers topics from text without requiring prior labels[5]. Unlike traditional bag-of-words models, BERT uses deep learning to learn semantic representations from large-scale text data, allowing it to capture semantic relationships between words more accurately and thereby enhance the expressiveness of the topic model. The steps for BERTopic modeling are shown in Figure 2.

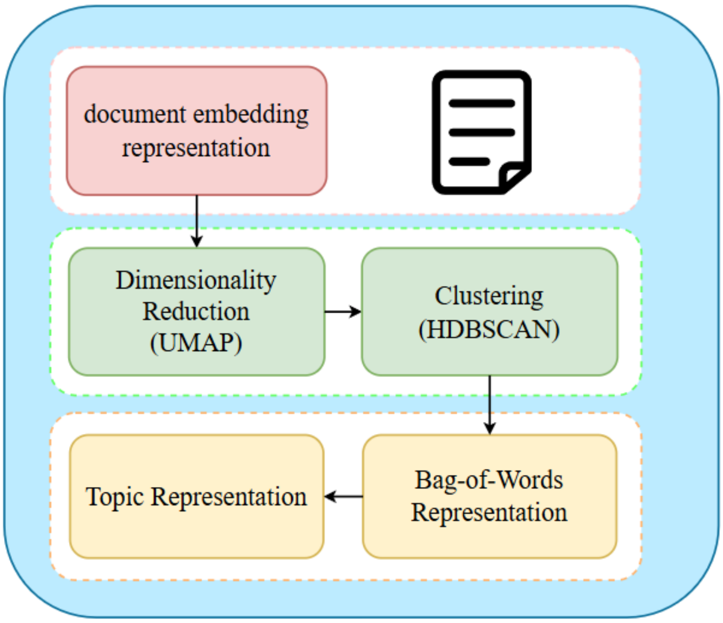
document embedding representation
Dimensionality Reduction (UMAP)
Clustering (HDBSCAN)
Bag-of-Words Representation
Topic Representation

Figure 2: BERTopic Topic Modeling Steps

The steps of the BERTopic model are as follows:

3.1.1. Document Embedding Representation

Document embedding representation involves mapping text data to a high-dimensional semantic space. In the BERTopic model, a pre-trained BERT model is used for document embedding representation. The embedding maps the text data into the semantic space learned by the BERT model, representing each document as a high-dimensional vector that preserves the semantic relationships and contextual information of the words within the document. This embedding representation allows the model to better understand the semantic content of the documents[6].

In this paper, we use Python's sentence_transformers library, where sentence_transformers is a Python framework developed based on Sentence-BERT to generate high-quality embedding vectors for sentences and short texts. Compared to the BERT model, sentence_transformers is simpler to use, and the vectors can be obtained directly by passing in the text.

3.1.2. Dimensionality Reduction

In the BERTopic model, a suitable dimensionality reduction algorithm can be chosen. This paper uses UMAP for dimensionality reduction. UMAP is a nonlinear dimensionality reduction algorithm that, compared to traditional linear methods like PCA, better preserves the nonlinear structure of the original high-dimensional data. UMAP involves two main steps:

(1) Learn the manifold structure in the high-dimensional space.

(2) Find the low-dimensional representation of this manifold.

3.1.3. Clustering

After dimensionality reduction, clustering is performed in the low-dimensional space using the HDBSCAN algorithm. HDBSCAN is an extension of DBSCAN that transforms DBSCAN into a hierarchical clustering algorithm, enabling it to discover clusters of varying densities. Consequently, HDBSCAN can handle clusters of different shapes and densities and is more robust to noise points[7].

HDBSCAN uses a soft clustering approach. Soft clustering allows the probability of a data point being assigned to multiple clusters rather than mandatorily assigning it to a defined cluster, which facilitates the handling of data points on boundaries, thereby increasing tolerance to noise and preventing irrelevant documents or data points from being assigned to any of the clusters. In addition, HDBSCAN treats noisy points as outliers. These outliers are not assigned to any cluster, which ensures that irrelevant data points do not affect the formation of any cluster.

3.1.4. Bag-of-Words Representation

To represent the subsequent topics, the probability of each word appearing in each cluster is required. The steps are as follows:

(1) Combine the documents in each cluster into a long document. For each cluster, merge all documents into one long document.

(2) Calculate the frequency of each word in the merged cluster document. The final result is a bag-of-words representation.

(3) Perform L1 normalization on the obtained bag-of-words representation to eliminate the effect of cluster size.

3.1.5. Topic Representation

During clustering with HDBSCAN, clusters are assumed to have varying densities and shapes. Traditional topic representation techniques, which are based on cluster centroids, can be inconsistent with this approach. To address this, the BERTopic model improves the topic representation method by using a category-based TF-IDF variant algorithm. The category-based TF-IDF variant algorithm is as follows:

(1) Calculate the frequency $tf_{x,c}$ of word x in the entire merged document c of the cluster. This step is obtained from the bag-of-words representation.

(2) Compute the inverse document frequency IDF_x, using the formula:

$$\mathrm{IDF_x} = \log(1 + \frac{\mathrm{A}}{\mathrm{f_x}}) \tag{1}$$

Where f_x represents the frequency of word x across all clusters, and A denotes the average number of words per cluster.

Step 3: Calculate the importance score $W_{x,c}$ of each word in each category, using the formula:

$$\mathrm{W_{x,c}} = \mathrm{tf_{x,c}} \cdot \mathrm{IDF_x} \tag{2}$$

3.2. Analysis of BERTopic Model Results

The BERTopic model does not require a predefined number of topics; it can automatically determine the final number of topics. By visualizing the inter-topic distances, we observe overlapping among multiple topics. From Figure 3, it is evident that the topics can be roughly divided into seven distinct areas:

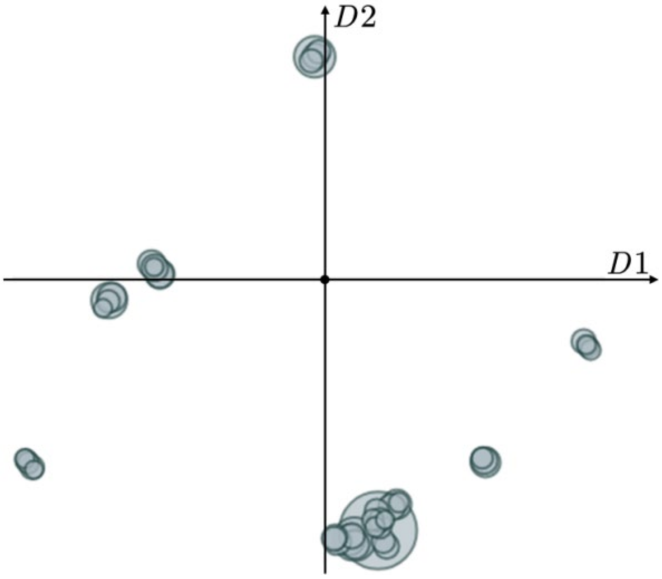

Figure 3: Inter-topic Distance Visualization

Thus, this paper selects representative topics from these seven areas for further analysis: Topic 0, Topic 2, Topic 3, Topic 4, Topic 5, Topic 6, and Topic 13. The topic word probability distributions and topic document histograms are shown in Figures 4 and 5, respectively:

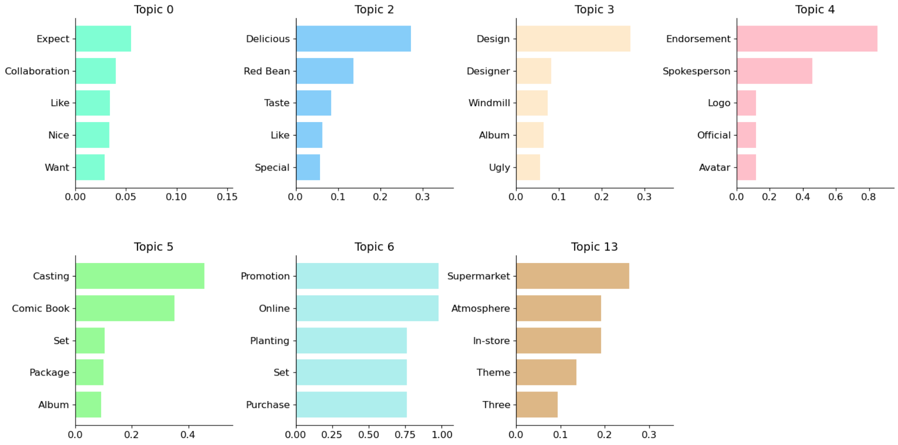

Figure 4: Topic Word Probability Distribution

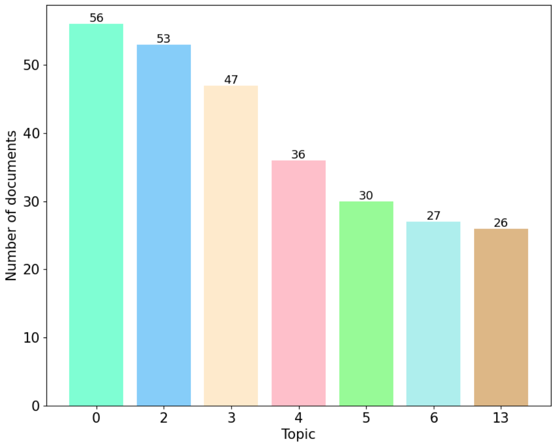

Figure 5: Number of Documents per Topic

From Figures 4 and 5, it can be observed that Topic 0 has the highest number of comments, with key terms including "expect", "collabration", and "want", indicating that consumers show a strong interest in co-branded products. Following that are Topic 2 and Topic 3, with a considerable number of comments, featuring terms such as "taste" and "want", which suggests that consumers mainly focus on the taste experience and packaging design of co-branded products. Topic 4 ranks third in the number of comments, with "endorsement" as a prominent keyword, highlighting that consumers choose to purchase products because of celebrity endorsements, which bring more attention and traffic to co-branded products. The final topics are Topic 0, Topic 5, Topic 6, and Topic 13. Topic 5 reflects consumers' interest in the free gifts included in joint packages, Topic 6 reflects consumers' recommendations of the products on promotional platforms, and Topic 13 reflects consumers' attention to offline themed stores, attracted by their unique ambiance.

4. Sentiment Analysis of Comments Using SnowNLP

Through the analysis of the above topic model, this paper has extracted themes from Weibo comments on the co-branded food economy, revealing consumer concerns under different themes. To gain a more comprehensive understanding of consumer sentiment, this paper will use the SnowNLP library in Python for sentiment analysis of the comments, further exploring consumer preferences for co-branded products[8].

4.1. Construction of the Sentiment Analysis Model

The SnowNLP library is based on the Naive Bayes algorithm. In this sentiment analysis process, each comment text under different themes will be classified into positive and negative sentiments[9].

The principle of Naive Bayes for sentiment analysis is based on Bayes' theorem, which describes updating the probability estimate for an event given new data, based on prior conditions. In sentiment analysis, we aim to estimate the probability that a text belongs to different sentiment categories based on the words in the text[10]. The basic principle of Naive Bayes sentiment analysis is as follows:

(1)Bayes' theorem expresses how to update the probability estimate for an event based on new observed data given prior conditions. Its mathematical expression is:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \tag{3}$$

In sentiment analysis, this paper represents the above formula as:

$$P(\mathrm{s}|\mathrm{t}) = \frac{P(\mathrm{t}|\mathrm{s}) \cdot P(\mathrm{s})}{P(\mathrm{t})} \tag{4}$$

In this context, $P(\mathrm{s}|\mathrm{t})$ represents the probability of a text belonging to a certain sentiment category given the text. $P(\mathrm{t}|\mathrm{s})$ denotes the probability of the text occurring given the sentiment category, $P(\mathrm{s})$ is the prior probability of the sentiment category, and $P(\mathrm{t})$ is the probability of the text occurring.

(2)Naive Bayes sentiment analysis is based on the assumption that each word in the text is conditionally independent. This assumption simplifies the computation. Where $P(\mathrm{w}|\mathrm{s})$ is the probability of the word w occurring given the sentiment category.

$$P(\mathrm{t}|\mathrm{s})=P(\mathrm{w}_1|\mathrm{s}) \cdot P(\mathrm{w}_2|\mathrm{s}) \cdot ... \cdot P(\mathrm{t}_n|\mathrm{s}) \cdot \tag{5}$$

Based on this assumption, this paper transforms the above probability expression into:

$$P(\mathrm{s}|\mathrm{t}) = \frac{\prod_{i=1}^{n} P(\mathrm{w}_i|\mathrm{s}) \cdot P(\mathrm{s})}{P(\mathrm{t})} \tag{6}$$

(3)To calculate the probabilities in Naive Bayes, parameter estimation is necessary. This paper uses the method of maximum likelihood estimation to solve the parameters, in order to estimate the probability of each word under different sentiment categories.

4.2. Solving and Analyzing Sentiment Analysis Results

Based on the results of the topic model analysis, this paper has distilled several core themes, including "taste experience," "packaging design," "celebrity endorsements," and "themed stores." These thematic directions reflect different points of focus in consumer comments. This section will use sentiment analysis to delve into the sentiment trends within these four themes.

For solving sentiment analysis results, this paper trains a model using a custom corpus, predicts sentiment categories for the collected comments, and calculates the sentiment proportions for each theme, which are then visualized in histograms.

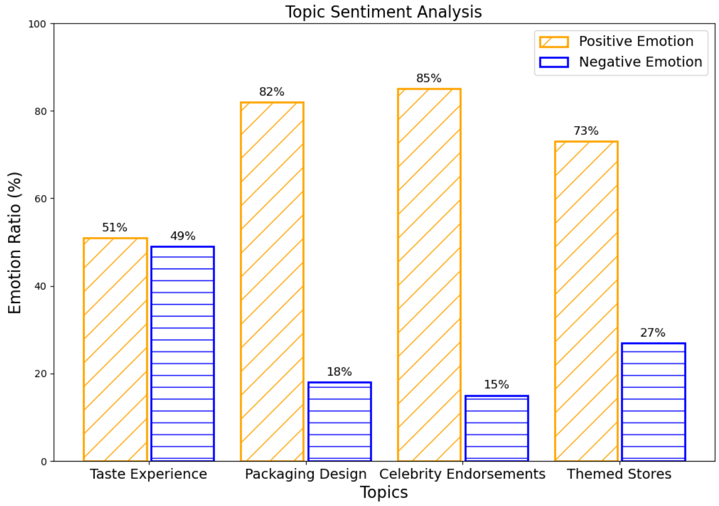

Figure 6: Sentiment Analysis Diagram

From Figure 6, it can be observed that the sentiment distribution for the "taste experience" theme is relatively balanced, with a small difference between positive and negative sentiments. Taking the co-branded product of Luckin Coffee and Moutai as an example, although the product incorporates some innovative flavor combinations, some consumers may not accept this novel taste. This suggests that while pursuing innovation is important, product design should also consider meeting the general taste preferences of the public.

In contrast, the sentiment distribution for the themes of "packaging design," "celebrity endorsements," and "themed stores" is more pronounced, with positive sentiments being dominant. In these themes, consumers are more likely to express strong positive emotions towards aspects such as product appearance, brand image, and offline event experiences. Celebrity endorsements bring more attention and appeal to the product, while the unique ambiance of themed stores creates distinctive experiences for consumers.

5. Conclusion

This paper explores user reviews of co-branded food products through crawling and analyzing comment data from Sina Weibo, utilizing natural language processing techniques such as BERTopic and SnowNLP. By extracting topics and performing sentiment analysis on the reviews, it is found that consumers' attention to co-branded products mainly focuses on celebrity endorsements, product packaging, taste experience, and offline activities.

Based on the research findings, this paper proposes the following recommendations from the perspective of businesses:

(1)Address Negative Taste Reviews

For products with negative taste reviews, brands should carefully consider consumer feedback, adjust flavors, or make product improvements to enhance user satisfaction. While pursuing taste innovation, it is also important to consider the general taste preferences of the public.

(2)Enhance Celebrity Endorsement Effectiveness

For products endorsed by celebrities, brands can further strengthen the association with the celebrity's image through more diverse marketing strategies to improve the effectiveness of the endorsement and meet users' expectations.

(3)Explore Cross-Industry Collaborations and Create Contrasts

Actively explore collaborations with different industries to integrate the advantages of various fields, which can help create intriguing and novel products. Introducing distinctly different elements to create contrasts in appearance and experience can attract consumer interest.

(4)Leverage Brand Effect

Collaborate with well-known, reputable, and high-image brands to fully utilize the influence of existing brands, using successful brands to drive traffic to new products.

(5)Focus on Product Packaging

Research the characteristics, preferences, and needs of the target consumer group in depth, focusing on designing unique packaging that aligns with the aesthetic tastes of the target audience.

(6)Promote Offline Themed Store Activities

Increase user participation in co-branded activities through offline events, enhance the fun of the activities, and strengthen the interaction between the brand and users.

References

[1] Rao J, Wang X. Analysis of brand crossover co-branding marketing strategies[C]. 2022 4th International Conference on Literature, Art and Human Development (ICLAHD 2022). Atlantis Press, 2023: 949-955.

[2] Wang M. Analysis of the Impact of Crossover Brand Co-branding on Consumer Purchasing Behavior[C]. Proceedings of the 6th International Conference on Economic Management and Green Development. Singapore: Springer Nature Singapore, 2023: 271-283.

[3] Yutong L. Hostility between men and women on Chinese social media: a content analysis of Sina Weibo with word frequency analysis and topic modeling. PsyArXiv, 11 Apr, 2023.

[4] Egger R, Yu J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts[J]. Frontiers in sociology, 2022, 7: 886498.

[5] GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794, 2022.

[6] Wang Z, Chen J, Chen J, et al. Identifying interdisciplinary topics and their evolution based on BERTopic[J]. Scientometrics, 2023: 1-26.

[7] Bhuvaneswari A, Kumudha M. Topic Modeling Based Clustering of Disaster Tweets Using BERTopic[C].2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon). IEEE, 2024: 1-6.

[8] Zhou B, Zhu Y, Mao X. Sentiment Analysis on Power Rationing Micro Blog Comments Based on SnowNLP-SVM-LDA Model[J]. Highlights in Science, Engineering and Technology, 2022, 4: 179-185.

[9] Hu N. Sentiment analysis of texts on public health emergencies based on social media data mining[J]. Computational and mathematical methods in medicine, 2022, 2022(1): 3964473.

[10] Xiong W, Zuo Y, Zhang M, et al. Research on Sentiment Analysis of E-commerce Live Comments based on Text Mining[J]. Frontiers in Computing and Intelligent Systems, 2023, 6(3): 34-36.