Journal of Research Administration
Society of Research Administrators International

CURRENT SITUATION OF CONSUMPTION CHARACTERISTICS OF THAI DURIAN PRODUCTS ON JD.COM PLATFORM IN CHINA

[1]Liurong Pan, [2]Boonsub Panichakarn
[1] [2]Faculty of Logistics and Digital Supply Chain,Naresuan University,65000 Phitsanulok,Thailand

Abstract— This paper takes consumers' comments from JD.COM as data to observe the consumption characteristics of Thai Durian in the Chinese market. Based on the Emotion Classification Model and Latent Dirichlet Allocation (LDA) Theme Model, characteristics of consumers' online shopping demand, characteristics of positive and negative emotional themes and the main influencing factors are discussed. Through an empirical research, it is found that: viewed from positive comments, consumers pay more attention to the freshness, taste, quality and benefits of the durian and the express logistics of the platform; seen from the negative comments, Internet consumers tend to focus more on the lack of fleshness, low pulp content and maturity. LDA Topic Model and topic mining visualization of positive reviews are used to figure out factors that push consumers to show positive emotional tendencies. These factors mainly conclude purchase experience, preference for durian products, shopping convenience and worthiness, durian taste and freshness. The factors leading to consumers' negative affective tendencies are after-sale problems, inconsistent quality of the durian, and invisibility of the inside of the durian.

Keywords: Thai durian, consumption characteristics, Emotion Classification, LDA Theme Model, Chinese market

I. INTRODUCTION

Online sales plays an increasingly important role in the retail industry as the size of the global e-commerce market continues to expand and the proportion of total retail e-commerce increases year by year. As such, to better decision-making and respond timely to customers' needs, more and more scholars and enterprises utilize computer technology such as text mining, machine learning and so on to locate information that consumers and businesses are more concerned about from a large number of reviews.

The positive effect of e-commerce on agricultural products, fresh products in particular, can be called "e-commerce enabling effect". By means of E-commerce, sales channels are expanded, distribution costs are greatly reduced and people's product awareness is enhanced. Therefore, e-commerce has been the impetus newly injected into the development of the agricultural industry. As a tropical fruit, durian is popular in the Chinese market as it is special in flavor. Thailand's durian exports to the Chinese market take up the largest share, with China importing 784,000 tons of fresh durian from Thailand, a deal of $3.848 billion, accounting for 95.04% and 95.34% of the total volume and total value respectively in 2022 (GACPRC, 2023)[1]. According to the Ministry of Commerce of Thailand, China was the largest exporter for Thai durian, taking up 96% of total exports, with a total export value of US$3.09 billion in 2022 (TMOC, 2023)[2]. Internet consumers favor such Thai durian varieties as Monthong, Chani and Kanyao.

Thai durian, with a huge market demand in China, faces competition from its equivalent in Indonesia, Vietnam, and Malaysia. Based on consumer reviews from e-commerce platforms, this study seeks to find out the needs of Chinese e-commerce consumers towards Thai durian products through sentiment analysis and LDA thematic research. It aims to to figure out factors that influence consumers' needs, and consumers' concerns about the product and their purchasing behavioral preferences. With this, it is much more easy for consumers to better understand product features of Thai durian and obtain advice of purchase. What's more, suggestions can be offered to e-commerce platforms and supplier merchants on marketing strategies for consumer demand to improve their competitiveness.

II. LITERATURE REVIEW

Online reviews provide information for consumers to know about products and merchants' after-sales services, affecting potential consumers' shopping habits and attitudes. These reviews also render data for merchants and platforms to understand consumers' needs[3]. For all this, consumers find it more difficult to locate information beneficial to their purchasing decision-makings while merchants find it an uneasy job to obtain information crucial to their business decision-makings as the reviews are in such a large number[4]. How to effectively utilize the reviews and provide feedback to consumers and merchants has become an urgent issue to be explored by all walks of life. In addition, with the rise of text mining technology, text mining offers processing solutions, of which the much more commonly-used are Sentiment Analysis and LDA Topic Model, etc. Computer technology makes it toilless for consumers and merchants to find information the two sides concerned about from a large number of reviews, reducing the workload of information extraction.

Data mining technology has been utilized in various fields in earlier researches[5].It is a recommendation system using data mining and preliminary association rules to analyze users' preferences on the Internet. Internet companies that know well consumer preference are better equipped to offer relevant electronic catalogs, to make product categories more attractive, and to encourage direct sales marketing through user recommendation systems. Kim and Chun (2019)[6]conducted a study to evaluate the merits and demerits of three competing automobile brands. They used Text Mining and Association Rules Techniques to analyze online reviews on cars. In a follow-up study.

Zhang and Raubal (2022)[7]proposed a framework that harnesses the Latent Dirichlet Allocation (LDA) technique to identify customer requirements. The framework consists of three distinct emotions - positive, negative, and neutral - used to analyze online reviews and extract significant product attributes. This research is highly relevant in shaping our comprehension of contemporary customers' expectations and preferences in the digital age. Majumder,Gupta,and Paul. (2022) [8]conducted a perceptive analysis of reviews on Amazon.com from three product categories: groceries, digital music, and video games since the e-commerce commenced. The researches mentioned above employed Text Mining techniques and uncovered Sentiment Polarization, identified sentiment patterns and assessed the perceived value of reviews.

Several studies have been conducted to investigate the impact of mass media on consumer sentiment. Blood and Phillips (1997)[9] discovered that economic news could affect consumer sentiment and lead to long-term economic implications. Gunther and Storey (2003)[10]proposed the Presumed Media Influence Model (PMIM). According to PMIM, individuals tend to change their attitudes or behaviors in the absence of direct information and the media has implications on public opinion. In the food industry, Tal-Or, Cohen, Tsfati and Gunther (2010) [11]used the PMIM to manipulate perceptions of news stories about the impending sugar shortages and to measure behavioral intentions by modifying the perceived exposure to the stories.

A recent study[12] was carried out to create online channels for purchasing fresh agricultural products. The study exploited the Latent Dirichlet Allocation (LDA) model to investigate the variables affecting consumer preference when purchasing online. Meanwhile,Meanwhile,Kusal, Patil,Choudrie,Kotecha,Vora,and Pappas(2022)[13] introduced a novel framework that synthesized customer opinions taken from product reviews. The framework effectively produced customized summaries by integrating deep neural networks, the LDA model, and grammar rules. Thus, it enables the user to derive meaningful insights from customer feedback by focusing on attributes of interest related to the product and on the most relevant comments.

Sentiment analysis, social network analysis, and topic modeling empower e-commerce businesses to gain a competitive advantage. These methods are mainly used to obtain detailed information about customer sentiment and product trends. The social media and Google Trends can be used by e-commerce businesses to gather operable competitive intelligence for data-driven decision-making. This study demonstrates how to apply these methods to a real-world case study and provides workable recommendations. E-commerce companies can use these tangible methods to extract competitive intelligence from data, to boost business growth and increase sales.

III. MATERIALS AND METHODS

IV. Data Collection

According to the statistics of Global E-commerce Network, China's e-commerce market is now mainly dominated by e-commerce platforms such as Alibaba, Pinduoduo, JD.COM, etc. This paper reflects the platform model, the frequency of user use and the platform evaluation structure and other factors, and ultimately selects the Thai durian on JD.COM platform as the object of study. This paper applies python software to write code to collect review data, and objectively selects the Thai durian products from 35 fresh fruit stores with top comprehensive rankings as research samples. The research process includes: (1) get the URL of each product; (2) set the headers, cookies, referrer and request header; (3) send request to the server. As the JD.COM review data is 'js' (JavaScript) dynamic page, it is necessary to carry out the 'json' format conversion, and save the obtained results as a 'csv' file and then the data collection is completed. The commodity data collected in this paper contains 38812 comments and it mainly includes user ID, evaluation time, comment score, and comment content, etc.. .

Table1. Sample Observations

Sampling period	19/02/2016-01/10/2023

Number of online platform stores	35
Information Collected	Username ID, review time, review score, review content
Sampled number of comments	38812
Default number of comments deleted	795
Number of duplicate reviews	7509

V. Conceptual Framework

The research was conducted from three aspects: the acquisition and collection of user comments from the e-commerce platform, the emotion classification analysis and LDA theme model analysis as shown in Figure 1. The data collection process includes the following steps: (1) select stores from JD.com; (2) search keywords “Thai durian”; (3) extract the text out; (4) pre-process the text; (5) analyze the emotion. Through machine learning, the classification experiment of the four classification models was conducted to complete the classification of positive and negative emotions. Going forward, the LDA Topic Model analysis was carried out. That means topic subdivision and influencing factor analysis were conducted. Finally, discussion was carried out and suggestions were made.

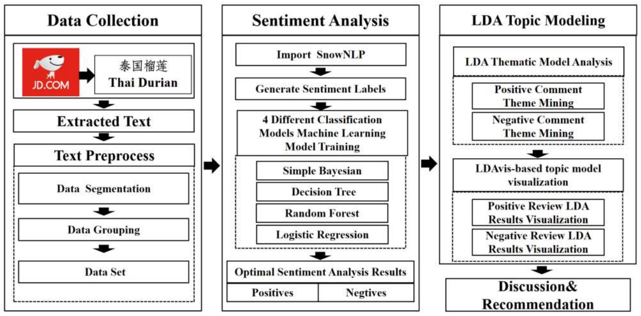
Data Collection
JD.COM
泰国榴莲
Thai Durian
Extracted Text
Text Preprocess
Data Segmentation
Data Grouping
Data Set
Sentiment Analysis
Import SnowNLP
Generate Sentiment Labels
4 Different Classification Models Machine Learning Model Training
Simple Bayesian
Decision Tree
Random Forest
Logistic Regression
Optimal Sentiment Analysis Results
Positives
Negtives
LDA Topic Modeling
LDA Thematic Model Analysis
Positive Comment Theme Mining
Negative Comment Theme Mining
LDAvis-based topic model visualization
Positive Review LDA Results Visualization
Negative Review LDA Results Visualization
Discussion& Recommendation

Figure.1 Methodology Framework

VI. Sentiment Analysis Techniques

VII. Naive Bayes Classifier (NBC)

Naive Bayes Classifier is a probability-based text classification method commonly used in sentiment analysis and other natural language processing tasks. Based on Bayes’ theorem, it assumes that the features in the text (usually words) are conditionally independent and it simplifies the probability calculations. In sentiment analysis, Naive Bayes Classifier is often used to categorize the text into positive, negative, or neutral sentiments.

Bayesian Learning Algorithm is the most practical learning approach for most learning problems and is based on evaluating explicit probabilities for hypotheses. NBC is extremely competitive with other learning algorithms and in many cases outperforms them. These algorithms are extremely important in machine learning since they provide unique perspectives for understanding many learning algorithms that do not explicitly manipulate probabilities. NBC is used in a variety of applications, including document classification (medical diagnosis systems performance management (14), and other fields.Domingos and Pazzani (1997) [14] prove optimality of the NBC under certain conditions even when the conditional independence assumption is violated.

The fundamental formula in NBC is based on Bayes' theorem:

$$P（B \mid A）= \frac{P(A \mid B)}{P(A)} \qquad (1)$$

When the naive Bayes method is used for classification, posterior probability is required and most of the time in a indirect manner. Therefore, the joint probability should be obtained through the prior probability and the conditional probability, and the there will emerge a reality that many instances will not satisfy the constraints and the probability may be 0. In the circumstances, the conditional independence assumption is applied to split the vector into features. The formula is expressed as follows:

$$\begin{aligned} &P（X=\mathrm{x} \mid Y=c_k)=P(X^{(1)},...X^{(\mathrm{n})}=\mathrm{x}^{(n)} \mid Y=c_k) \\ &=\prod\nolimits_{j=1}^{n} P(X^{(1)}=\mathrm{x}^{(1)} \mid Y=\mathrm{c_k}) \qquad (2) \end{aligned}$$

After the joint probability is obtained from the probabilistic model, the classifier is constructed by learning the maximum posterior Probability P (Y=ck | X=x). The the largest class of posterior probability is taken as the output of x class. The formula is expressed as follows:

$$P(Y=c_k \mid X=\mathrm{x})=\frac{P(X=\mathrm{x} \mid Y=c_k)P(Y=c_k)}{\sum_k P(X=\mathrm{x} \mid Y=c_k)P(Y=c_k)} \qquad (3)$$

The naive Bayes classifier is the classification when the posterior probability is maximum, which is expressed as the formula:

$$\mathrm{y}=\arg\max c_k P(Y=\mathrm{c_k})\prod\nolimits_{\mathrm{j}} p（X_{（\mathrm{j}）}=\mathrm{x}_{(\mathrm{j})} \mid Y=\mathrm{c_k}） \qquad (4)$$

VIII. Decision Tree

Decision tree is a logical approach to data mining. It is widely used in information extraction, prediction and classification. The algorithm is easy to learn in the process of data information learning even if the learner has only a little background knowledge. Decision tree induction approach provides insights on conditional relationships between independent variables and a target variable. It also facilitates abduction, deduction, and induction processes, enabling researchers to postulate hypotheses (abduction) based on empirical observations and to statistically test them (induction) to generate a theoretical model [15]. It consists of decision nodes, branch nodes and leaf nodes. The topmost leaf in the decision tree serves as the root node. Each branch is a new decision node, and each leaf node represents a possible categorical attribute. The process proceeds as follows:

First, the information entropy of the category attributes is calculated. Entropy is employed to measure the impurity or randomness of a data set .[16] It is assumed that the data set S is a collection of n samples of testing attributes, and the testing attributes are C_i(i= 1.2.... .n), then the total information entropy of category attributes is：

$$entropy(S) = -\sum_{i=1}^{n} P_i \ \log_2 P_i \qquad (5)$$

Pi denotes the probability of the testing attribute Ci.

Second, the information gain of testing attributes is calculated. It is assumed that the testing attribute C_i={a1,a2... am} divides S into m copies, S_j denotes the j_{th} subset after C_j divides S, and |S|and $|S_{ij}|$ denote the number of samples in the set respectively, then the information entropy of testing attribute Ci is expressed as ：

$$entropy(S, C_i) = \sum_{j=1}^{m} \frac{|S_{ij}|}{|S|} entropy(S_{ij}) \qquad (6)$$

The information gain of attribute C_i to S is ：

$$gain(S, C_i) = entropy(S) - entropu(S, C_i) \qquad (7)$$

Finally, the testing attribute with the largest information gain is selected as the branch node of the decision tree. So on and so forth, each branch node is used as a new decision node to continue the above process, and finally the decision tree is constructed.

IX. Random Forest

Variable selection methods for random forest classification are thoroughly described in the literature [17],[18]. Random Forest is a powerful ensemble learning model widely employed in sentiment analysis and text classification. It comprises multiple decision trees and enhances model diversity through random sampling and feature selection. The model expression formula is as follow:

$$f_{RF}(x) = Major(h(x, a_i)) \qquad (8)$$

Here h (x, a;) is the prediction result of a single decision tree; f_{RF} (x) is the category with the most frequent prediction, namely the output of the random forest.

Multiple training datasets are created by random sampling, with each used to build a decision tree. Feature subsets are randomly chosen as split criteria. Each decision tree provides a prediction, and the final result is either a majority vote (for classification) or an average (for regression). It is a collection of decision trees, each with its own set of rules. The final prediction results from aggregating the opinions of multiple trees to improve model performance. Random Forest excels at handling text data and accurately predicting sentiment labels.

X. Logistic Regression

Logistic regression is actually an extension of linear regression [19]. Logistic regression is a universal model in sentiment analysis for such binary classification tasks as determining whether sentiment is positive or negative. It models the relationship between text features and sentiment labels.

The formula of logistic regression model is as follows:

$$P\ (y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \qquad (9)$$

Here, P(y=1) represents the probability of predicting a positive sentiment, $(x_1,x_2,...,x_n)$ are text features, and $(\beta_0,\beta_1,\beta_2,...,\beta_n)$ are model parameters.

Logistic regression is suitable for modeling the relationship between text features and sentiment, making it useful for such sentiment classification tasks as determining whether sentiment is positive or negative.

XI. Latent Dirichlet Allocation (LDA) Modeling

Topic modeling belongs to unsupervised learning that relies on statistical models mostly in machine learning. LDA topic modeling is an unsupervised machine learning algorithm that discovers implicit textual patterns with a three-layer Bayesian probability of words, topics, and documents as its core structure.

The LDA model assumes that each review consists of a random mix of topics in a certain proportion and that the mix obeys a multinomial distribution. It is denoted as：

$$Z \mid \theta = Multionomial(\theta) \tag{10}$$

Each theme consists of a mixture of words from the vocabulary list in a certain proportion, and the proportion of the mixture also obeys the multinomial distribution and it is recorded as：

$$W \mid Z, \varphi = Mulinomial(\varphi) \tag{11}$$

The probability of producing the word w under comment d_j is denoted as:

$$P\text{（}\mathrm{w}_{\mathrm{j}} \mid d_j) = \sum_{s=1}^{k} P(w_i \mid z = \mathrm{s}) \times P(z = s \mid d_{\mathrm{j}}) \tag{12}$$

$P(w_i \mid z = \mathrm{s})$ denotes the probability that the word belongs to the s|th theme, and $P(z = s \mid d_j)$ denotes the probability of the s|th theme under the comment d_j.

The approximate estimation of parameters θ and φ by the LDA model is usually done by using Gibbs sampling, a special case of the Markov Chain Monte Carlo (MCMC) algorithm. Parameter estimation of the LDA model using Gibbs sampling is based on the following equation:

$$P\text{（}z_i = s \mid Z_{-i}, W) \propto \text{（}\mathrm{n}_{\mathrm{s,-i}} + \beta_i)/(\sum_{i=1}^{v} n_{s,-i} + \beta_i) \times (n_{s,-i} + \alpha_s) \tag{13}$$

The formula z_i =s| labels the probability that the word w^i belongs to the s|th topic. The word Z_{-i} denotes the probability of all other words, $n_{s,-i}$ denotes the number of words that are assigned to the current topic z_s except the current word w_i , and $n_{s,-j}$, denotes the number of documents that are assigned to the current topic z_s except the current document d_j.

Through the derivation of the above equation, it can be derived that the parameter estimates $\varphi_{s,i}$ for the word w_i in topic z_s, and the multinomial distribution $\theta_{j,s}$ for of topic z_s in comment d_j . The formulas are as follows：

$$\varphi_{s,i} = (n_{s,i} + \beta) / \sum_{i=1}^{v} n_{s,i} + \beta_i) \tag{14}$$

$$\theta_{j,s} = (n_{j,s} + \alpha_s) / \sum_{s=1}^{k} n_{j,s} + \alpha_s) \tag{15}$$

Here $n_{s,i}$ denotes the number of occurrences of the word w_i in the topic z_s, and $n_{j,s}$ denotes the number of topics z_s contained in the document d_j.

XII. Data Pre-processing

The data pre-processing contains the following steps. First, the stop word list file namely the HITU Stop Word List is opened by means of the open function in the Python programming language and read by UTF-8 encoding. This file contains some common stop words. Second, the list of stop words is initialized. The list contains a number of generic stop words that typically appear in textual data but do not provide actual semantic information. These generic stop words include “durian”, “customer service”, “Jingdong”, etc. Third, the processes of segmentation and stop words removal are complete. A regular expression (re.compile('\d+')) in the code was used to define a pattern (pattern) to match numbers. All this was to recognize and filter numbers in the subsequent processing. Fourth, the code initializes an empty list value to store the processed text. The code then enters the main loop of data processing and iterates through each line of the text data named “Data Evaluation Content”. Throughout each line, the jieba.lcut (line) function of jieba library is used to perform a disambiguation, breaking the sentence into individual words that were stored in the list Segs. Fifth, an empty list words is initialized by the code to store the processed words. In the inner loop, each word in the Seg is traversed for further processing. The words are checked to see if they were equal to '\r\n', and if so, they will be skipped to avoid processing blank characters. Regular expression patterns is used to check if the words contain numbers and stop words. The final step is to retain and remove common stop words to finalize the processed text data.

XIII. RESULT AND DISCUSSION

XIV. Sentiment Characteristics Analysis

SnowNLP, a sentiment analysis tool was applied to perform sentiment analysis on text data and generate sentiment labels. The core idea of the method is to determine the sentiment tendency of a text based on its sentiment score. The first check ensures that the text is non-empty. Then, the text is sentiment analyzed by the SnowNLP library to assign a sentiment score to the text data. If the sentiment score is greater than or equal to 0.1, the text is labeled as positive sentiment (1), indicating that the text has a positive sentiment; if the sentiment score is less than 0.001, the text is labeled as negative sentiment (0), meaning that the text has a negative sentiment. NaN (Not-a-Number) is returned for texts of length 0 or other exceptions. In this particular sentiment categorization task, category 1 (called “positive sentiment”) contained 25,601 samples while category 0 (called “negative sentiment”) possessed only 4,907samples.

The review data was analyzed in a continuous manner for a more objective analysis. Four different classification models (Bayesian, Decision Tree, Random Forest, and Logistic Regression) were used to find more accurate results of sentiment analysis. As to Natural Language Processing (NLP) and Machine Learning tasks, pre-processing and feature extraction of textual data are crucial to transforming textual data into a digital form that can be processed by machine learning models. At the same time, partitioning the data into training and testing sets contributes to the evaluation of the model performance.

XV. Text Feature Extraction and Data Segmentation

The dropna () function was first used to remove the rows containing NaN (missing values) from the data to ensure data integrity. Next, two columns were selected from the data, merged and labeled. The merge column contains the text data and the label column possesses the corresponding labels. Then, the label column was converted to an integer type for subsequent classification tasks.

Going forward, the train _test_ split function was used to implement data splitting, where the parameter test_size=0.2 specifies that 20% of the data is used as the test set and the remaining 80% is used as the training set. The parameter random _state=42 was used to set the random seed to ensure consistent data splitting results for each run.

Finally, a TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer, i.e., TfidfVectorizer was used as the feature extraction method. This vectorizer transformed the text data into word-based feature vectors (where the parameter max_features=1000 specifies the maximum number of features to be 1000) . This helps to reduce the dimensionality and increase the computational efficiency. In addition, smooth_idf=True and use_idf=True were set to enable smooth IDF (Inverse Document Frequency) computation to improve the weight representation of features.

XVI. Model Training

As to text categorization tasks, choosing the right model is critical to the final performance. For each model, a 10-fold cross-validation was performed and several performance metrics were evaluated, including Accuracy, Precision, Recall, F1 Score (F1), and Area Under the ROC Curve(AUC).

Table 2 shows the training results of each model:

Table 2. Classification Evaluation Metric

Evaluation indicators	Accuracy	Precision	Recall	F1	AUC
Bayesian value	0.54	0.99	0.51	0.68	0.73
Decision Tree value	0.96	0.97	0.99	0.98	0.73
Random Forest value	0.96	0.97	1.00	0.98	0.96
Logistic Regression value	0.97	0.97	1.00	0.99	0.98

It can be seen from the above results that Decision Tree, Random Forest and Logistic Regression models perform better with high accuracy, precision, recall and F1 score. Random

Forest and Logistic Regression models in particular perform well in several performance metrics and are of relatively high AUC values of 0.96 and 0.98 respectively.

Therefore, the training results speak for Random Forest and Logistic Regression models when the optimal models are selected. The final choice depends on the specific task requirements and optimization goals. If higher recall and F1 score are needed, Logistic Regression model is to be chosen; if higher AUC value is needed, Random Forest model is to be chosen. Model selection should fit specific application scenarios and trade-offs in performance metrics.

In the performance evaluation of the logistic regression model, the confusion matrix provides detailed information about the classification results. Based on the confusion matrix provided, ,conclusions can be drew as shown in Figure 2:

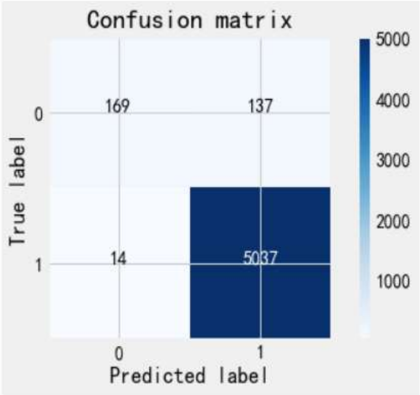

Figure 2 Visualization of the Model Confusion Matrix

True Positives (TP): the logistic regression model succeeded in correctly predicting 169 positive category samples as positive categories. These samples were accurately identified as belonging to the target category.

False Positives (FP): the model mispredicted 137 negative category samples as positive categories. These samples actually belonged to the negative category, but were incorrectly categorized as positive by the model.

True Negatives (TN): the model predicted exactly 5037 negative category samples as negative categories. These samples were accurately identified as not belonging to the target category.

False Negatives (FN): The model predicted incorrectly 14 positive category samples as negative. These samples actually belonged to the positive category but were incorrectly categorized as negative by the model.

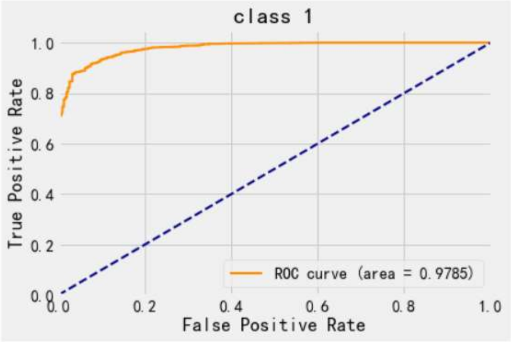

Figure 3 roc Plot

In addition, as shown in Figure3 , the Area Under the Curve (AUC) value of 0.9785 indicates the area is under the Receiver Operating Characteristic (ROC) curve of the model. AUC is a metric used for evaluating the classification performance of a model and is commonly used for binary classification problems. Specifically, an AUC value of 0.9785 implies that the model excels in classifying positive and negative categories. The value 0.9785 is close to 1, indicating that the model performs well under different thresholds.

XVII. Consumer Sentiment Characteristics Visualization

The size and color of words in the graphic reflect the frequency and importance. Based on the results of sentiment classification in the previous chapter, this paper uses the word cloud library in Python to make word cloud maps for positive and negative reviews. After importing the positive and negative reviews, elements that are displayed include the background image selected, the number of high-frequency words, style such as font color, and the number of LDA themes. After that, word cloud maps are drew. Displaying the word cloud maps can visualize the consumer's concerns and the characteristics of durian shown in the durian reviews. The positive and negative word cloud diagrams are shown in Figures 4.

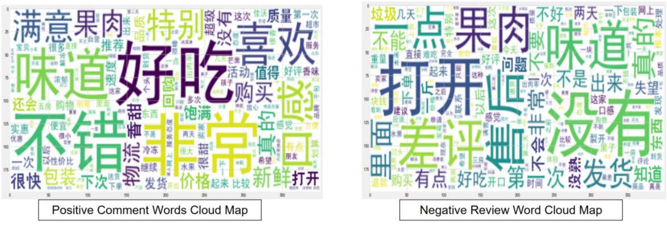

Figure 4 Comment Words Cloud Map

"It can be seen from the positive word clouds of durian that the words that appear more frequently are "good", "delicious", "very", "flavor", "like", "special", "taste", "satisfaction", "flesh", "full", "sweet", "fresh", "very fast", "packaging", "logistics", "quality", etc. It can be seen that consumers are more concerned about the freshness, taste, quality, cost-effectiveness of durian and the platform's express logistics. The fact that consumers concern highly about express logistics indicates that consumers have a positive attitude towards the above characteristics of Thai durian sold on JD.COM. The concern is also an important factor affecting consumers' positive sentiments.

It can be seen from the negative word clouds of durian that the frequency is high in such words as "bad evaluation", "no", "garbage", "unpalatable", "won't", "after-sales service", "under-ripe”, “disappointment”, “a little”, “problem”, and “return”. The high frequency reflects that Thai durian appears to lack pulp inside after consumers receive them. In addition, words such as “after-sales”, “return” show consumers’ dissatisfaction. Negative evaluation words appear including “under-ripe”, and it reflects that consumers doubt about the maturity. The negative words above symbolize that consumers are in negative emotion towards the Thai durian sold on JD.COM. However, implications of the word “poor” and “bad” are vague, so it is impossible to tell what aspect is bad for consumers. Therefore, it is necessary to further analyze the theme model to see under which theme these words appear most frequently.

XVIII. Consumer Characteristics LDA Thematic Model

LDA topic clustering via Python modeling is based on word-topic probability matrix. The whole process contained several steps. First, the number of LDA themes were determined . Before the theme analysis of the reviews, the number of themes should be determined to avoid overlap and confusion, and to get a better theme differentiation. Positive sentiment is categorized into four themes and negative sentiment into three themes. Then, LDA theme result is analyzed. After determining the optimal number of themes in the previous section, the python software is used to perform theme mining for the positive and negative reviews on Thai durian and to analyze each theme. Mining results of reviews on positive (positive) themes are obtained according to the analysis above. The results are shown in Figure 5.

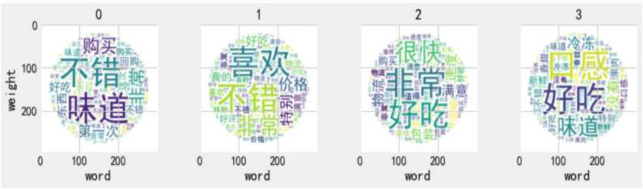

Figure 5 Positive Comment Theme Mining

It can be seen from Figure 5 that there are different theme words under each theme and that the characteristic words under some themes do not fit well with the overall tendency of the theme, but this hardly impacts the overall theme judgment.

The results of the four themes are analyzed on the basis of positive feature words of durian reviews in the table:

The high-frequency feature words in Theme 0 are “nice”, “flavor”, “purchase”, “very”, “first time”, “tasty”, “something”, “repurchase”, and “this time”. These words suggest that the theme may be related to shopping experience, including the evaluation of products, the frequency of shopping and the degree of satisfaction.

The high-frequency feature words in Theme 1 are “nice”, “like”, “plentiful”, “price”, “special”, “tasty”, “really”, “logistics”, and “satisfied”. These words suggest that this theme may involve product preferences and evaluations of product features, as well as evaluations of logistics and service related to the products.

The high-frequency feature words of Theme 2 are “very”, “tasty”, “very fast”, “satisfied”, “logistics”, “convenience”, “packaging”, “purchase”, and “worthwhile”. These words reflect that this theme may be relevant to the shopping convenience and the value of shopping, namely the swiftness of shopping, the satisfaction with the goods, the speed of logistics and packaging evaluation.

The High frequency feature words of Theme 3 are “tasty”, “texture”, “flavor”, “frozen”, “fresh”, “sweet”, “pulp”, “special”. These words indicate that the theme may involve the evaluation of the taste, texture and freshness of the product. These words contain the evaluation of the quality of the product as well as descriptions of specific products.

It can be concluded from the above that negative comments are categorized into three themes, as shown in Figure 6.

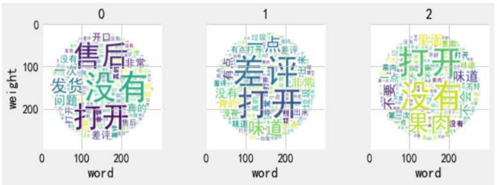

Figure. 6 Negative Comment Theme Mining

The results of the four themes are analyzed on the basis of negative feature words of durian reviews in the table:

High-frequency feature words of Theme 0 are “no”, “open”, “after-sale”, “delivery”, “problem”, “once”, “non-profit”, “really”, “poor”, and “opening”. These words show that the theme may involve negative emotions related to product purchase and after-sales service and reflect dissatisfaction and problems in the shopping process.

High-frequency feature words of Theme 1 are “bad review”, “open”, “taste”, “a little”, “very”, “no”, “a little”, “really”, “junk”, “out”. These words indicate that the theme may be pertinent to poor reviews of product quality and characteristics. These keywords contain negative evaluation and description concerning the product experience. These words may imply that the user is not satisfied with the product’s performance.

High-frequency feature words of Theme 2 are “no”, “open”, “pulp”, “flavor”, “inside”, “don’t”, “won’t”, “a little”, “bad”. These words imply that the theme may involve problems of and dissatisfaction within the product. These keyword disclose possible problems and user dissatisfaction. These words may reflect negative evaluations of product quality and experience.

XIX. LDA Visual-based Topic Model Visualization

After theme mining of positive and negative comments, the pyLDAvis package in python is used to visualize the theme results. The visualization of the results can verify the accuracy of the number of themes in the previous section. With this, differentiation between themes can be seen more intuitively.

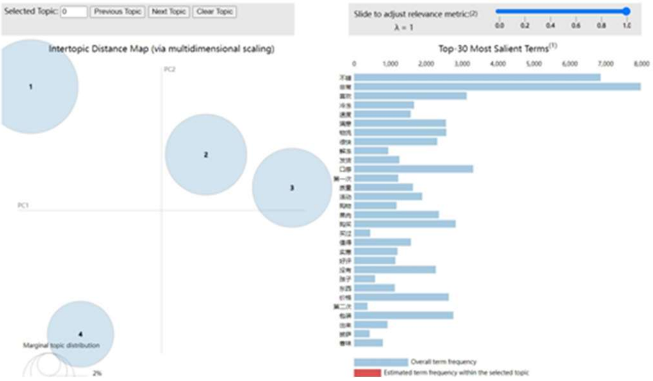

Figure 7 Positive Review LDA Results Visualization

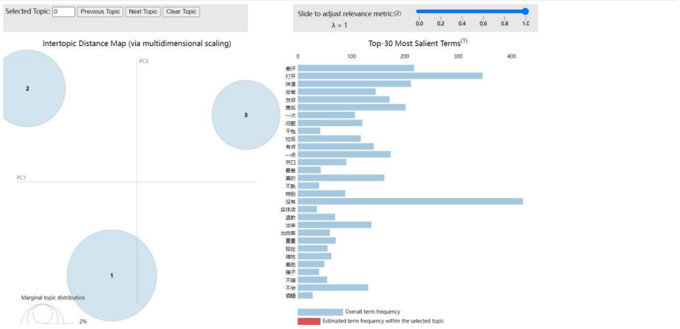

Figure 8 Negative Review LDA Results Visualization

Circles in Figure 7 and Figure 8 do not overlap in the two-dimensional spatial quadrant, and the size of each circle represents the number of words under each theme. From Figure 7 and Figure 8, it can be seen that there is a clear spatial difference between the four themes of the positive comments and the three themes of the negative comments. The distribution of each individual theme is more dispersed, which verifies the accuracy of the theme categorization above. As such, a conclusion can be drew that the themes have been better differentiated from each other.

The main factors prompting consumers to generate positive emotions are durian shopping experience, durian product preference, shopping convenience and worthiness, durian taste and freshness. The main factors pushing consumers to show negative emotions are after-sale problems, durian quality, and durian internal problems as shown in Table 3.

Table 3 Different Emotional Influences on Consumers

Emotion	Main influencing factors	Subject keywords
Positive emotions	Shopping experience	Nice, purchase, very, first, repurchase
	Durian product preference	Favorite, rich and often, special, tasty, really
	Shopping convenience and product value	Very fast, satisfied, logistics, speed, worth it
	Durian taste and freshness	Frozen, texture, flavor, freshness,

		sweetness
Negative emotions	After-sales service problems	After-sales, delivery, problems
	Quality of durian	Bad reviews, smell, garbage
	Problems after opening the durian	Flesh, inside, not, a little, bad.

Conclusion

Based on the positive and negative review data of Thai durian consumers from JD.COM e-commerce platform, word cloud maps were created by selecting words with high word frequency to show consumers' emotional characteristics. LDA Topic Model and visualization for topic mining of positive review texts were used to figure out factors prompting consumers to show positive emotional tendencies. These factors mainly conclude purchase experience, preference for durian products, shopping convenience and worthiness, durian taste and freshness. The factors leading to consumers' negative affective tendencies are after-sale problems, inconsistent quality of the durian, and invisibility of the inside of the durian. The results of the research are informative to stakeholders to better their decision makings to better fulfill the consumers' needs. Be that as it may, there are still much room for improvement in this study. In the future research, data source obtained from multiple e-commerce platforms will be studied to conduct comparative analysis among multiple platforms in this regard.

ACKNOWLEDGEMENTS

This research is supported by the Faculty of Logistics and Digital Supply Chain, Naresuan University, Thailand. to grant a Ph.D. scholarship to the first author. There is no specific grant from any funding agency in the public, commercial, or non-profit sectors.

REFERENCES

[1] General Administration of Customs of the People's Republic of China(GACPRC), (2023). Import and Export Trade Statistics Database, http://stats.customs.gov.cn/

[2] Thailand's Ministry of Commerce （ TMOC ） (2023) Fruit Trade Data, https://www.moc.go.th/

[3] I. Ventre, and D.Kolbe, The impact of perceived usefulness of online reviews, trust and perceived risk on online purchase intention in emerging markets: A Mexican perspective. Journal of International Consumer Marketing, 32(4),287-299,2020.doi:10.1080/08961530.2020.1712293

[4] F.Sudirjo, F.Ratnawati, R.Hadiyati, I. N. T.Sutaguna, and M. Yusuf,THE INFLUENCE OF ONLINE CUSTOMER REVIEWS AND E-SERVICE QUALITY ON BUYING DECISIONS IN ELECTRONIC COMMERCE. Journal of Management and Creative Business, 1(2), 156-181,2023.doi:10.30640/jmcbus.v1i2.941

[5] S. H.,Liao,R.Widowati,and Y. C. Hsieh, Investigating online social media users' behaviors for social commerce recommendations. Technology in Society, 66, 101655,2021. doi:10.1016/j.techsoc.2021.101655

[6] E. G.Kim,and S. H.Chun, Analyzing online car reviews using text mining. Sustainability, 11(6), 1611, 2019.doi:10.3390/su11061611

[7] Y.Zhang,and M. Raubal,Street-level traffic flow and context sensing analysis through semanti c integration of multisource geospatial data. Transactions in GIS, 26(8), 3330-3348, 2022.doi:10.1111/tgis.13005

[8] M. G.Majumder,S. D.Gupta,and J. Paul,Perceived usefulness of online customer reviews: A r eview mining approach using machine learning & exploratory data analysis. Journal of Business Research, 150, 147-164, 2022. doi:10.1016/j.jbusres.2022.06.012

[9] D. J.Blood,and P. C.Phillips, AEconomic Headline News on the Agenda: New Approaches t o Understanding Causes and Effects.@ In Communication and Democracy: Exploring the Intelle ctual Frontiers in Agenda-Setting Theory, edited by Maxwell McCombs, Donald L. Shaw, and D avid Weaver, 97-113,1997.doi:10.4324/9780203810880

[10] A. C.Gunther, and J. D.Storey, The influence of presumed influence. Journal of Communication, 53(2), 199-215,2003.doi:.1460-2466.2003.tb02586.x

[11] N.Tal-Or, J.Cohen, Y.Tsfati and A. C. Gunther, Testing causal direction in the influence of presumed media influence. Communication Research, 37(6), 801-824,2010. doi:10.1177/0093650210362684

[12]C.Xie, X.Tian, X. Feng, X.Zhang,and J.Ruana, Preference Characteristics on Consumers' On line Consumption of Fresh Agricultural Products under the Outbreak of COVID-19: An Analysis of Online Review Data Based on LDA Model. Procedia Computer Science, 207, 4486-4495,2022.doi:10.1016/j.procs.2022.09.512

[13]S. Kusal, S.Patil, J.Choudrie, K. Kotecha, D.Vora,and I.Pappas, A Review on Text-Based Emotion Detection--Techniques, Applications, Datasets, and Future Directions. arXiv preprint arXiv:2205.03235, 2022.doi:10.48550/arXiv.2205.03235

[14] P.Domingos, and M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29, 103-130,1997.doi:10.1023/A:1007413511361

[15] K. M.Osei-Bryson,and O.Ngwenyama, Using decision tree modelling to support Peircian ab duction in IS research: a systematic approach for generating and evaluating hypotheses for syste matic theory development. Information Systems Journal, 21(5), 407-440,2011.doi:10.1111/j.1365-2575.2010.00368.x

[16]R. Molala, Entropy, Information Gain, Gini Index—The Crux of a Decision Tree. Medium,2020.https://blog.clairvoyantsoft.com/entropy-information-gain-and-giniindex-the-crux-of-a-decision-tree-99d0cdc699f4.

[17] A.Hapfelmeier, and K.Ulm,A new variable selection approach using random forests. Compu tational Statistics & Data Analysis, 60, 50–69,2013.doi:10.1016/j.csda.2012.09.020

[18] L.N.Sanchez-Pinto, L. R.Venable, J. Fahrenbach, and M. M.Churpek, Comparison of variable selection methods for clinical predictive modeling. International Journal of Medical Informatics, 116, 10–17, 2018.doi:10.1016/j.ijmedinf.2018.05.006

[19] P.Schober,and T. R. Vetter, Logistic regression in medical research. Anesthesia and analgesi a, 132(2), 365,2021.doi:10.1213%2FANE.0000000000005247