Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240564

What affects customers' online shopping behavior, research that applied machine learning to Amazon product reviews

Ziqin Qin[1,*,†], Kefan Dong[2,4,†], Bingzhao Xie[3,5,†]

[1]Adam Smith Business School, The University of Glasgow, Glasgow, G128QQ, UK
[2]School of Information, Shanxi University of Finance and Economics, Taiyuan, 030012, China
[3]School of Business, University of Connecticut, Storrs, CT, 06269, United States

[*]m15773629585@163.com
[4]dkf595915002@gmail.com
[5]xavierxie08@gmail.com
[†]These authors contributed equally to this work and should be considered co-first authors.

Abstract. The application of Natural Language Processing (NLP) in marketing has undergone significant evolution, with machine learning algorithms playing a crucial role in extracting valuable insights from complex textual data. This study focuses on comparing the performance of Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and a specialized sentiment analysis model, Latent Dirichlet Allocation (LDA), in the context of online platform reviews. While previous research has delved into individual algorithms, there is a paucity of horizontal comparisons. Suitable algorithms for sentiment analysis on online platform reviews, specifically for Amazon, were filtered in this work. A dataset from Kaggle (https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews) comprising 500,000 reviews and 10 columns was utilized, overcoming time and resource constraints by opting for secondary data analysis. The primary objective was to assess the performance metrics of SVM, RF, NB, and LDA in classifying reviews into positive, neutral, and negative sentiments. Despite the massive size of the dataset posing challenges to the accuracy of the algorithms, nuanced results in precision, recall, and F-score were observed, not replicated in prior studies. Attempts to enhance accuracy by switching vectorizers yielded marginal improvements. Interestingly, LDA emerged as a transformative model, leveraging its ability to generate WordClouds for a systematic analysis of customers' emotional attachments. In addition to sentiment analysis, an investigation into the identification of factors influencing consumer purchasing behavior on Amazon was conducted. By training the LDA model on positive, neutral, and negative comments, distinctive features associated with each sentiment category were extracted. This analysis aims to unravel the underlying product features that contribute significantly to customer decision-making processes. In conclusion, this work provides a comprehensive evaluation of SVM, RF, NB, and LDA in the realm of sentiment analysis on Amazon product reviews. The findings shed light on the challenges posed by large datasets, the limitations of traditional vectorizers, and the unique capabilities of LDA in uncovering emotional nuances. Moreover, the investigation into consumer purchasing behavior offers valuable insights for marketers seeking to understand the factors influencing online shopping decisions.

Keywords: Machine learning, customer reviews, sentiment analysis.

DOI: 10.54254/2755-2721/76/20240564

1. Introduction

In the ever-evolving landscape of e-commerce, Amazon stands as a behemoth, attracting millions of consumers worldwide. However, what shapes the shopping experience of these Amazon consumers? To answer this question, this research embarks on a journey into the intricate world of Amazon's e-commerce platform and the wealth of product reviews it hosts. Leveraging the power of machine learning, we delve into the intricate web of consumer behavior, uncovering the specific factors that influence the Amazon shopping experience. This exploration promises to shed light on the profound impact of product reviews and other variables, ultimately providing valuable insights for both consumers and sellers in the Amazon marketplace.

2. Literature Review of Current Academic Research

2.1. Classifying sentiment of customer reviews

Since the rise of e-commerce, traditional offline stores have been replaced due to the overwhelming advantage that completely changed consumers' choices during shopping. Between consumers and sellers, the centuries-old traditional relationship was already broken down. In the e-commerce platform, different goods become fair in price, transparent and convenient to purchase, customers can evaluate different aspects of products before purchasing. What's more, instead of relying on the information provided by sellers, online reviews can provide valuable details of products, services, and shopping experiences. This kind of review is more likely to gain the trust of potential consumers since people believe they are more reliable. At the same time, when consumers lack knowledge about products or fail to form valid expectations about consumption results, they will actively seek other information to reduce the risk of uncertainty [1]. In this new form of relationship, sellers need to be extra careful about online reviews to ensure that they will not influence potential customers implicitly. The literature studied the impact of online reviews and coupons on product sales and prices, and their research results are consistent with the belief of people that positive reviews have a positive impact on product sales. However, it is worth noting that negative reviews have even greater impacts on subsequent sales than positive reviews, and the negative impact is more pronounced than positive reviews. While it is pretty easy to judge people's attitudes towards a product by their rating, the problem becomes more complicated when it comes to customer reviews. Because reviews differ in length and language, it is difficult to use a simple method to judge people's attitudes.

In this context, scholars tried to predict the sentiment of reviews by applying machine learning techniques and natural language processing [2,3]. Some scholars judge the review based on the corresponding product, with higher ratings meaning the review is positive (e.g., 4, 5 on a 5-point scale) and lower ratings meaning the review is negative (e.g., 1, 2 on a 5-point scale) [4].

However, some scholars have noted the huge impact of negative reviews on people's shopping behavior and have expanded beyond simply categorizing reviews into 3 categories. Chaffar and Inkpen used machine learning algorithms and identified six emotions: disgust, anger, happiness, fear, surprise, and sadness. And they think that two kinds of emotions, disgust and anger, should be paid more attention in the emotion analysis studies [5]. Because in today's world, people's disgust and anger can strongly affect others' buying behavior. In the meantime, online product evaluations use advanced techniques to analyze consumer thoughts and preferences, including categorizing reviews as positive, mixed, neutral and negative, as well as composite scoring methods. A notable finding was the increase in the number of online reviews after 2008, which may be related to the financial crisis and increased online shopping. A study employed TextBlob and NLTK techniques to process the data and used TF-IDK for accurate analysis. This study emphasized on subjective and emotional features of consumer reviews and quantitative analysis through polarity and subjectivity [6].

Yang and several academics strongly agree with Chaffar and Inkpen's conclusion on emotional dictionaries and point out that, "The core of the sentiment lexicon-based approach is to construct a sentiment lexicon [5,7]." In this context, they used a sentiment lexicon to find the emotions of the text. The sentiment lexicon is constructed by selecting appropriate sentimental words, degree adverbs,

DOI: 10.54254/2755-2721/76/20240564

negative words, sentimental intensity and sentimental polarity are marked for the constructed sentiment lexicon. After the text has been input, the words in the text will be matched with the sentiment words in the sentiment lexicon, and the matched words are weighted and summed to obtain the sentiment value of the input text, thereby determining the sentimental polarity of the input text according to the sentiment value [8]. Therefore, this study builds on a specific sentiment dictionary based on the results of machine learning to analyze the core value needs of customers. And this dictionary can provide insight for this study.

2.2. Common Model used in sentiment analysis studies

Besides the diverse sentiment classifications, scholars also applied different machine learning algorithms for sentiment analysis. Kaushik and their group primarily focus on the analysis of Amazon reviews, aiming to perform sentiment classification using various machine learning approaches, including Logistic Regression, Random Forest, Naïve Bayes, and Bidirectional Long-Short Term Memory [1]. Haque and his group recognize the significance of product reviews in the digitalized e-commerce landscape, aiming to simplify the decision-making process for customers by polarizing reviews and utilizing classifier models like Naive Bayes, Logistic Regression, and Random Forest to predict sentiment orientation on new data [8]. Their result suggests that BERT emerged as the top performer, achieving an impressive accuracy rate of 94.7%. Additionally, the Bi-directional Short-Term Memory model delivered a respectable accuracy of 93%.

Geetha and Renuka mainly focused on identifying useful information based on customer reviews to analyze customers' interest preferences. In their study, ABSA (Aspect Sentimental Analysis) is used as a base to analyze the reviews of two major categories of electronic products from Amazon. Algorithms such as SVM, Naïve Bayes Classifier, and LSTM are used to classify customer reviews. The BERT (Bidirectional Encoder Representations from Transformers) model is used for natural language processing, in the final experiment, the BERT model is significantly better than other models and has higher accuracy [9].

On the other hand, Yang and his groups also combined the advantages of sentiment lexicon, CNN model, GRU model and attention mechanism to propose the “SLCABG” model [7]. They used the book review data collected on Dangdang websites in China to compare different sentiment analysis models. However, limitations have been marked, this model can only divide emotions into positive and negative categories, which is not suitable for the field with high requirements for emotion refinement.

Another literature that is highly like this study has been flagged, with Gani and Tomimatsu conducting a sentiment study on Amazon game customer review data. They also proposed a model, like what Yang had proposed that divides customer reviews into positive and negative user emotions—The phenomenal-granularity Language Model (GLMP) method. Unlike Yang and his framework, this model provides mood levels, increments, acceleration, and recommendations. In terms of the latitude of the model, GLMP has a remarkable contribution to the increase and acceleration of consumer sentiment level, but this particular change varies with the variable of time. Therefore, this model is more suitable for emotion analysis with time variables. On the other hand, like the SLCABG model, the GLMP model is also limited by the sentiment dichotomy and cannot conduct a more detailed classification of consumer sentiment [7,10].

3. Methodology

In this sentiment analysis of customer reviews, we leverage a comprehensive dataset sourced from Amazon product reviews, including customer ratings and reviews. This dataset serves as a valuable foundation for understanding customer sentiments. The general framework of this study is shown in Figure 1.

DOI: 10.54254/2755-2721/76/20240564

Figure 1. Proposed Framework.

3.1. Data Preprocessing

3.1.1. Pre-processing of the current dataset. To have usable data, the extracted data will be extracted and filtered and subsequent analysis will be carried out on this basis. Data preprocessing is done by filtering the text comments in the selected dataset by length, followed by extracting 10,000 data randomly from each score interval, and then finally removing special characters (such as
 #$&) from the filtered text comments and standardizing the text to lowercase to get a clean dataset at the end. The processed results will be stored in a new CSV file for subsequent calls.

3.1.2. Stop word processing. When processing Amazon product review data, we adopted a method of removing stop words to improve the quality and accuracy of the text. We introduced a collection of English stop words through the NLTK library to effectively remove these common and uninformative words from the review text. The function first splits the text into words and retains only the words that are not in the set of stop words through a list comprehension. Finally, the filtered words are recombined through spaces to obtain the comment text processed by stop words. The main goal of this step is to reduce the noise in the text and extract key information that is meaningful for tasks such as sentiment analysis.

3.1.3. Stemming. To further refine the review text, we introduced a stemming method, using the PorterStemmer stemmer of the NLTK library function. This function splits the text into words and converts each word into its base form to reduce vocabulary diversity. By applying a stemmer to each word in the review text, we achieve the goal of mapping different variations of vocabulary to the same concept. Finally, by recombining the processed words with spaces, we get review text with more consistent and standardized expressions. This step helps improve the model generalization ability for subsequent natural language processing tasks, analyzing review data more accurately and meaningfully.

3.2. Vectorizer

In the implementation of the vec_fun function, we use CountVectorizer, a classic text vectorization method. CountVectorizer builds a word frequency matrix for each text based on simple word frequency statistics. By setting the ngram_range parameter, we can explore the multi-level structure of the text with different phrase lengths, allowing the model to better understand the text context. The output of CountVectorizer is a matrix containing all words in the text and their occurrence times, which provides the model with a rich feature set to more comprehensively reflect the linguistic characteristics of the review. On the other hand, to consider the importance of words in the entire text collection, we introduce TfidfVectorizer. This method combines Term Frequency (TF) and Inverse Document Frequency (IDF) to reduce the impact of common words by giving higher weight to words that are high in frequency but appear less frequently in the entire text collection. TfidfVectorizer also supports processing phrases of different lengths by setting the ngram_range parameter. Similar to CountVectorizer, the output of TfidfVectorizer is a sparse matrix, where each row corresponds to a text sample and each column

DOI: 10.54254/2755-2721/76/20240564

corresponds to a feature, representing a word in the text. By introducing TF-IDF weights, we more accurately capture the importance and contribution of words in the review text.

3.2.1. Parameter Adjustment. In the process of building a machine learning model, parameter adjustment is an important step in optimizing model performance. This study adopted two different parameter adjustment methods, namely Chi-square and Exhaustive Parametric Grid Search CV.

3.2.2. Chi-square parameter adjustment. The Chi-square test is a commonly used feature selection method to select features that have a significant impact on the target variable by evaluating the independence between features and labels. We use the Chi_square_fun function to perform feature selection on Amazon review data before building the model. This function selects the k features with the highest correlation with the target variable using the chi-square test, where k is specified by the parameter k_in. This step helps eliminate features that contribute little to the model and improves the model's computational efficiency and generalization ability.

3.2.3. Exhaustive Parametric parameter adjustment. To find the optimal combination of model hyperparameters, we employed exhaustive parameter grid search cross-validation. In the model_fun function, we use GridSearchCV to search for model parameters by setting different hyperparameter combinations. Specifically, we set different parameter combinations for different classifiers (Random Forest, Support Vector Machine, Naive Bayes), such as the number and depth of trees in the random forest, the C value and kernel function type of the support vector machine, the smoothing parameters of Naive Bayes, etc. This step enables the model to conduct a comprehensive search within a given parameter space to find the optimal configuration, improving the model's performance and generalization ability.

3.3. Model Selection

3.3.1. Naïve Bayes(NB). Naïve Bayes is a simple classifier based on bayes theorem. NBs classifier assumes that each input is independent and has the same importance. While the assumption of NB can not be satisfied in real-world contexts, it is widely applied in sentiment analysis since it’s the advantage of using low computational effort [11].

3.3.2. Supportive Vector Machine(SVM). A supervised learning method is used for categorization. SVM aims to find a hyperplane or a group of optimal hyperplanes that separates different classes in a dataset. This model can maximize the margin between data points of different classes, leading to lower error [12].

3.3.3. Random Forest (RF). An ensemble learning method based on decision trees. It builds multiple decision trees during training and merges their predictions for more robust and accurate results. By combining the predictions of multiple trees, RF shows more resilience to overfitting and exhibits improved generalization performance.

3.3.4. Latent Dirichlet Allocation (LDA). Except for the three models introduced earlier in this chapter, The Latent Dirichlet Allocation (LDA) model is also brought to help develop customers’ emotional research. In the beginning, academics developed the LDA model just to study population inheritance related to biology [13]. Blei, Ng and Jordan these three people first discovered the impressive performance of LDA models in machine learning. With an excellent understanding of de Finetti's theorem, Blei, Ng and Jordan hope that the LDA model can capture important intra-document statistical structures through mixed distributions [14]. The basic logic is to represent a document as a random mixture of potential topics, each of the topics contained is characterized by a word distribution.

DOI: 10.54254/2755-2721/76/20240564

In other words, LDA solves the problem of analyzing which topics are present in a given article and the proportion of each topic that appears, so the result of model fitting will present the core keywords and specific probabilities for each topic. It assumes modelling that each topic is a mixture of certain underlying words, and each document is a mixture of topic probabilities. Worth noticing, LDA is an unsupervised generative probability method for corpora

In addition, Jelodar et al. argued that building an LDA model requires a large data set. The minimum necessary size depends on the characteristics and average length of the document. Due to the increase in observation values, the larger the data set is, the better the result gets. In this study, 50,000 comments were used for sentiment analysis, which is very suitable for the dataset requirements of the LDA model [15].

LDA assumes that every document can be represented as a probability distribution of potential topics and that the topic distribution in all documents has a common prior of Dirichlet. Figure 2. Given a corpus composed of M documents, one of the documents inside M is called "I" and it contains Ni words ($i \in 1,... M$), LDA modelled the corpus according to the following generation process:

1. Choose a multinomial distribution φ (k) for topic K($k \in \{1,..., K\}$) from a Dirichlet distribution with parameter β

2. Choose a multinomial distribution $\theta(i)$ for the document i($i \in \{1,..., M\}$) from a Dirichlet distribution with parameter α.

3. For a word Wz ($z \in \{1,..., N\}$) in document i, Select a topic Zi from θi, Select a word Wi from φzi

In the above generation process, the words in the document are only observed variables, the others are latent variables (φ and θ) and hyperparameters (α and β).

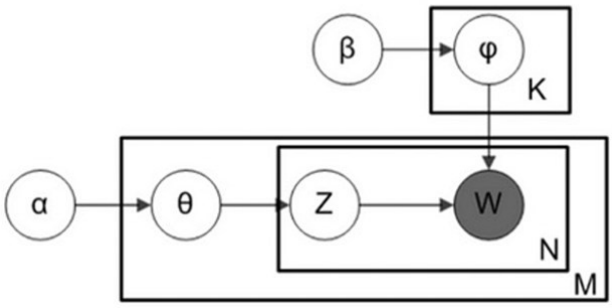

Figure 2. Probabilistic graphical representation of the LDA model [15].

The Dirichlet distribution allows sampling of a probability distribution on a probability simplex where all numbers add up to 1 and these numbers represent K different classes of probabilities. The K-dimensional Dirichlet distribution is an open standard $(K - 1)$-simplex, it has a k parameter and is expressed as the uncertainty of probability distribution [15]. In addition, K is the number of topics, M is the number of all documents used in this analysis, and N is the number of words in the document. The Dirichlet- multinomial pair for the corpus-level topic distributions, is considered as (α, θ). φ is the word distribution for the topic K, and θ is the topic distribution for the document i. The variables ZMN and WMN are word-level variables that are sampled once for each word in each document. Finally, the Alpha parameter is the Dirichlet previous concentration parameter. Dirichlet previous concentration parameter. The previous Dirichlet concentration parameter is mainly used to represent the topic density of the document, and as the alpha value increases, the document is assigned to more topics for classification, resulting in a more specific topic distribution for each document. Beta represents the same prior concentration parameter for the word density of the topic, and at higher beta values, the model assumes that the topic is composed of most words of the same type and results in a more specific word distribution for each topic [14].

DOI: 10.54254/2755-2721/76/20240564

3.4. Model Training and Testing

To gain insights into the factors shaping customers' purchasing decisions in the realm of online shopping, we leveraged machine learning models on Amazon product reviews. The model training phase involved the extraction of pertinent features, such as pre-processed text and sentiment labels, from the reviews. Employing techniques like TF-IDF vectorization and Chi-square feature selection, we prepared the data for diverse machine learning models, including Random Forest, Support Vector Machine (SVM), and Naive Bayes. A crucial step in this process was the hyperparameter tuning performed through grid search for Random Forest, SVM, and Naive Bayes, optimizing their configurations for better performance. Subsequently, the chosen model, dictated by the user through the model_name parameter, underwent training on the pre-processed dataset. For Naive Bayes, the trained model was saved using joblib. dump for future use.

Following the model training, the testing phase aimed to assess the models' efficacy in predicting consumer sentiments and, by extension, their purchasing decisions. The dataset was split into training and testing sets to ensure unbiased evaluation. The trained models—whether Random Forest, SVM, or Naive Bayes—were then applied to predict sentiments on the testing data. Evaluation metrics such as precision, recall, and F-score were calculated to gauge each model's ability to capture the intricacies of customers' decision-making. In the case of Naive Bayes, the pre-trained model was loaded using joblib.load before making predictions on the testing data. Additionally, the parameters of the loaded model were printed, shedding light on its configuration. This comprehensive approach aimed to unravel the nuanced elements affecting customers' buying decisions during online shopping through the application of machine learning techniques to Amazon product reviews. Before implementing the LDA model, this study clarified the use of Python libraries. Currently, the relatively common libraries involving the LDA model include Sci-Kit Learn(sklearn) and Gensim. This study adopts a combined approach. Specifically, in the main part of the research, sklearn packages are imported and used for analysis, mainly because sklearn is clearer in data analysis than Gensim. The visualization work of this study is all realized by Gensim, mainly because Gensim has more advantages in visualization.

At the same time, fine-tuning hyperparameters is considered necessary. For example, this study focuses on the LDA model's topic number (K), feature number (V), Perplexity matric, topic consistency, and Dirichlet prior parameters of alpha and beta. For the LDA model, the number of topics is one of the most important hyperparameters because it has only one input variable. The selection of the number of topics mainly resulted from the size and characteristics of the data set. For example, a topic analysis for a given item may require fewer topics than a multi-category topic analysis. The larger the dataset, the greater the number of topics, and only if the dataset represents a diverse set. In this study, two indexes of theme, Perplexity matric and Theme consistency were used to select the number of topics. The visualization software package pyLDAvis was introduced in Python to help this study better understand and explain the relationship between each topic.

3.5. Topic perplexity tolerance and topic consistency

In this study, topic perplexity and topic consistency are two important indicators used to evaluate the number of topics. Before beginning, it is necessary to mention the relevant concepts of perplexity. The Perplexity metric is one of the most important indicators for the evaluation of language models. Perplexity explains the degree to which models are surprised by captured data that models have not seen before. This study chose 1-15 as the desirable range for the number of topics, mainly due to the large dataset size (50,000 comments). The visualization of perplexity showed a major inflexion point when the number of topics was 4—Figure 3. Therefore, for this study, 4 is the optimal number of topics based on the perplexity metric.

DOI: 10.54254/2755-2721/76/20240564

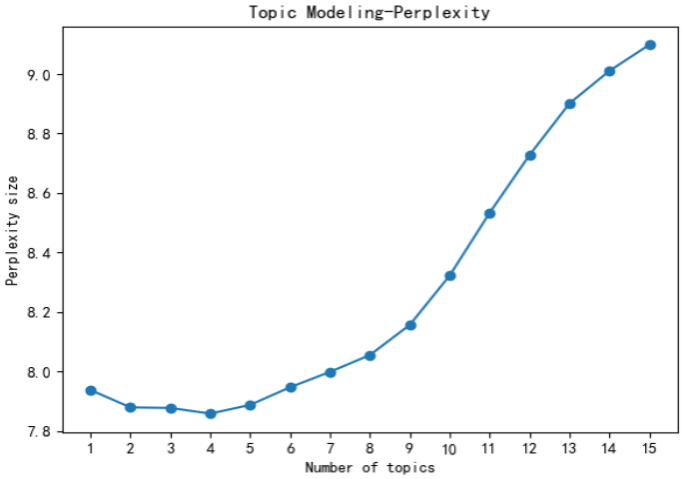

Figure 3. Topic Modeling Perplexity.

However, recent studies have pointed out that the pursuit of perplexity may result in results inconsistent with human judgment, or even opposite outcomes [16]. Therefore, this study also introduced topic consistency to help researchers determine the optimal topic number parameters as well. Topic consistency measures the score of each topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements will help distinguish semantically interpretable topics from those that are the product of machine learning and statistical inference. Unlike the perplexity metric, conformance sets specific limits on machine learning based on human-understandable semantics. As shown in Figure 4, the number of topics reaches the best consistency at 11. However, when the number of themes exceeds 6, the room for increasing consistency becomes smaller.

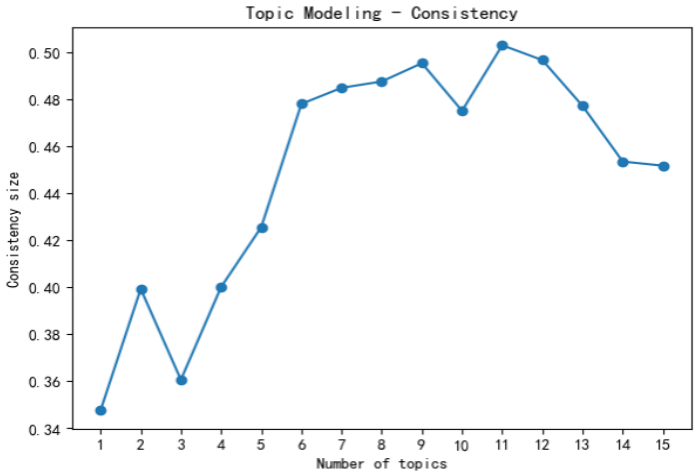

Figure 4. Topic modelling Consistency.

When the number of topics is 6, the perplexity value is low, which indicates that the model has a good performance in fitting data. This may mean that topic number _6 is a relatively better choice because it better captures the underlying structure and topic correlation of the data. On the other hand, perplexity increases significantly when the number of topics exceeds 6, which may indicate that the model is overfitting the data or is too sensitive to noise. In this case, the use of 6 topics may be more explanatory and practical.

DOI: 10.54254/2755-2721/76/20240564

3.6. Alpha and beta Dirichlet prior parameters

As Dirichlet's previous concentration parameters, alpha and beta represent the density of the document to topic, and the density of topic to word, respectively. A high alpha value usually means the document is assumed to contain more topics. However, as alpha decreases, sparsity increases. When sampling the distribution, the alpha value is zero or close to zero in most cases. Therefore, in this study, having an alpha value of doc_topic_prior=0.1 is considered appropriate. Beta parameters will formulate the subject's prior beliefs about word sparsity and consistency, adjusting for biases that certain subjects will favour certain words. Higher beta values result in a larger and more complex mix of data for each topic. And the lower the beta is, the lower the linguistic richness of the words within the subject. However, for a dataset with a sufficiently symmetrical distribution, a beta value of 0.01 is sufficient.

3.7. Future numbers

In this study, tf-idf vectorizer value ranking was adopted. In general, the number of features is determined by the size of the data set and the number of topic clusters. However, the number of functions affects the training time of the LDA model. In addition, the number of features is also limited by the device hardware. Therefore, the number of 1000 features is taken as the hyperparameter to experiment.

4. Results

4.1. Model Evaluation

The performance—Precision, Recall, and F-score—of three machine learning models are evaluated (see Table 1). Two different vectorization techniques, CountVectorizer and TfidfVectorizer, are employed. Of all the models, the RF model exhibited the highest precision (0.723) and high recall (0.665) when using TfidfVectorizer, showcasing its efficacy in capturing both positive and negative sentiments. Conversely, the CountVectorizer yielded varied results across models. Specifically, Naive Bayes demonstrated low precision (0.656) and relatively lower recall (0.662). In addition, using TfidfVectorizer positively impacted the performance of the SVM, with higher precision (0.688) and recall (0.705) compared to its CountVectorizer counterpart.

Table 1. Performance of different models and vectorizers.

Model	Vectorizer	Precision	Recall	F-score
NB	CountVectorizer	.656	.662	.656
RF	CountVectorizer	.704	.657	.601
SVM	CountVectorizer	.646	.651	.647
RF	TfidVectorizer	.723	.665	.605
SVM	TfidVectorizer	.688	.705	.687

4.1.1. Feature Importance. As shown in Figures 7 and 8, the feature importance in RF contains more words relating to negative emotion(bad, worst, disappoint), which can not be seen in the feature importance of NB. The feature importance of these two models both contain common positive vocabulary. *This* provides some insight into the categorization of different machine-learning models.

DOI: 10.54254/2755-2721/76/20240564

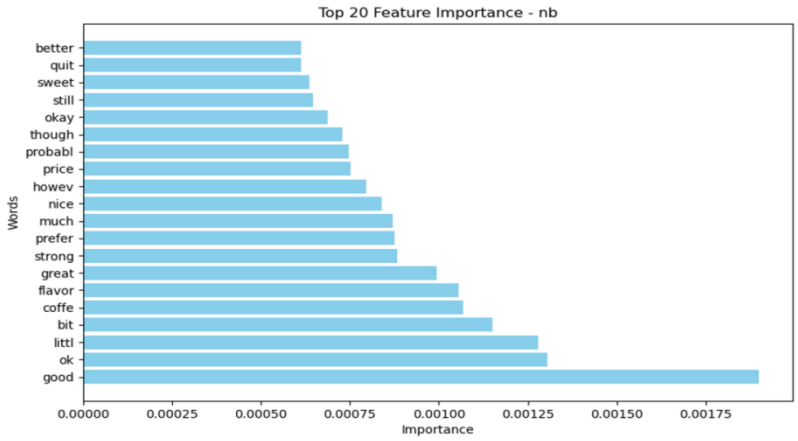

Figure 5. Feature importance of Naive Bayes (NB).

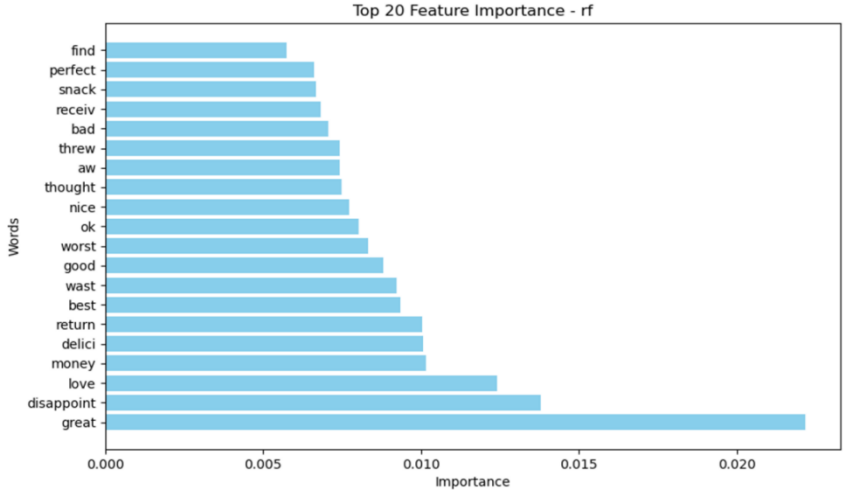

Figure 6. Feature importance of Random Forest (RF).

4.2. LDA Results Analysis

4.2.1. Theme 1: Product Flavor and Texture. As shown in Figure 9. In examining factors influencing customer online shopping decisions, we employed machine learning techniques to theme model Amazon product reviews. The results from the LDA (Latent Dirichlet Allocation) topic analysis revealed several decisive factors. Theme 1 primarily focuses on the flavor and texture of products, accounting for 21.3% of the dataset's vocabulary. Key terms within this theme include 'taste,' 'like,' 'sugar,' 'flavor,' 'chocolate,' and 'good.' The frequency of these terms is not only significantly higher in Theme 1 compared to the overall dataset average but also prominently positioned in red bars in visualizations, indicating a high customer focus on food taste and texture in product evaluations. Apart from taste descriptors, Theme 1 also encompasses terms related to the product itself, such as 'bar,' 'snack,' 'cereal,' and 'milk,' closely linked to food products. The occurrence of the word 'healthy' suggests that health factors are also a crucial consideration for customers when selecting food items. Furthermore, terms like 'sweet,' 'water,' 'fruit,' and 'coconut' may indicate customer diversity in taste preferences and specific flavor likings. The in-depth analysis of Amazon review data under Theme 1 demonstrates that product flavor and texture are among the key factors affecting customer purchase decisions online. This finding is vital for online retailers as it underscores the importance of highlighting taste and texture in product descriptions and marketing.

DOI: 10.54254/2755-2721/76/20240564

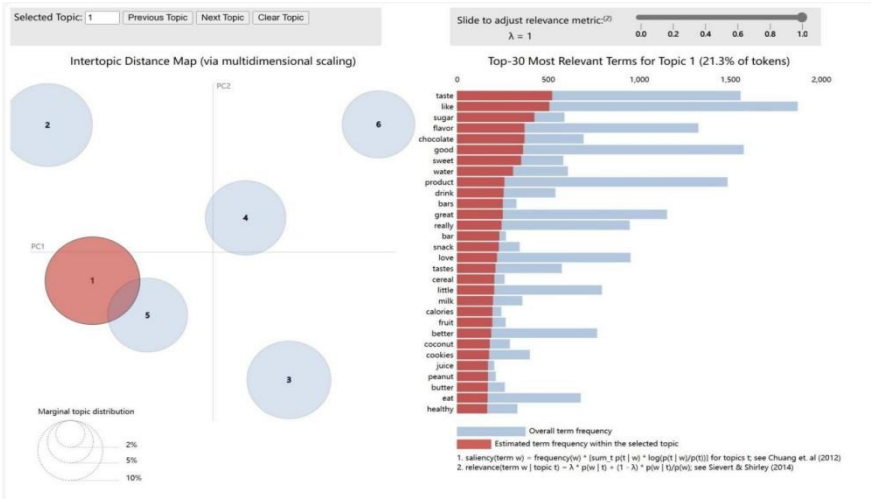

Figure 7. The LDA Model Themes Three one: flavor and texture of products.

4.2.2. Theme 2: Coffee and Beverage Choices. As shown in Figure 10. The machine learning analysis further revealed another key factor influencing customer online shopping decisions: the selection of coffee and other beverages. Theme 2, through LDA topic analysis, covers 19.1% of the dataset's vocabulary, highlighting specific terms related to beverages. In Theme 2, the term 'coffee' significantly outweighs other terms in frequency, indicating its status as one of the most discussed products in online reviews. Concurrently, terms like 'tea,' 'cup,' 'like,' 'flavor,' and 'taste' suggest that customers focus not only on product type but also on flavor and sensory experience when evaluating products. Additional terms such as 'strong,' 'bitter,' 'roast,' 'weak,' 'green,' 'vanilla,' and 'bold' reflect varying customer preferences for coffee strength and flavor. The term 'keurig' likely refers to a specific coffee machine brand or product, and the appearance of 'pods' suggests a popular coffee preparation method. The analysis of Theme 2 indicates that the flavor, preparation method, and brand of beverages like coffee and tea are significant considerations in the customer decision-making process. This insight is valuable for retailers, highlighting the need to focus on these attributes in online product descriptions and promotions.

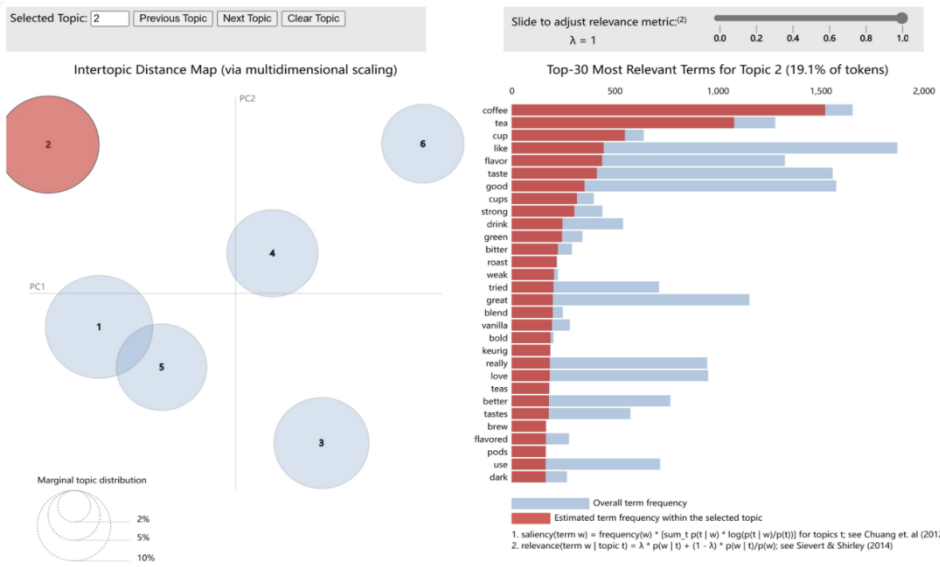

Figure 8. The LDA Model Themes Two: the selection of coffee and other beverages.

DOI: 10.54254/2755-2721/76/20240564

4.2.3. Theme 3: Pet Supplies and Care. As shown in Figure 11. In analyzing elements influencing customer online shopping decisions, Theme 3 focuses on pet supplies, encompassing 16.7% of the dataset's vocabulary. This theme's vocabulary pertains to pet food and care, highlighting customer concern for pet health and happiness. Terms such as 'dog,' 'food,' 'treats,' 'cat,' and 'chew' point to customer preferences and considerations in selecting pet food. Additionally, 'loves,' 'good,' 'baby,' and 'great' express customer emotional evaluation and satisfaction with the products. The occurrence of 'ingredients,' 'smell,' 'china,' 'small,' and 'bag' in this theme may relate to the product's origin, packaging size, and food quality, reflecting customer consideration not just for pet preferences but also for product safety and practicality. 'Formula,' 'treat,' and 'eat' suggest high customer attention to product formulation and pet dietary health in purchasing pet supplies. LDA Theme 3 analysis shows that online shopping comments about pet supplies often revolve around pet dietary preferences, product quality, and expressions of brand trust. This guides online pet product sales strategies, emphasizing the importance of clearly marking ingredients, emphasizing product safety, and showcasing customer reviews.

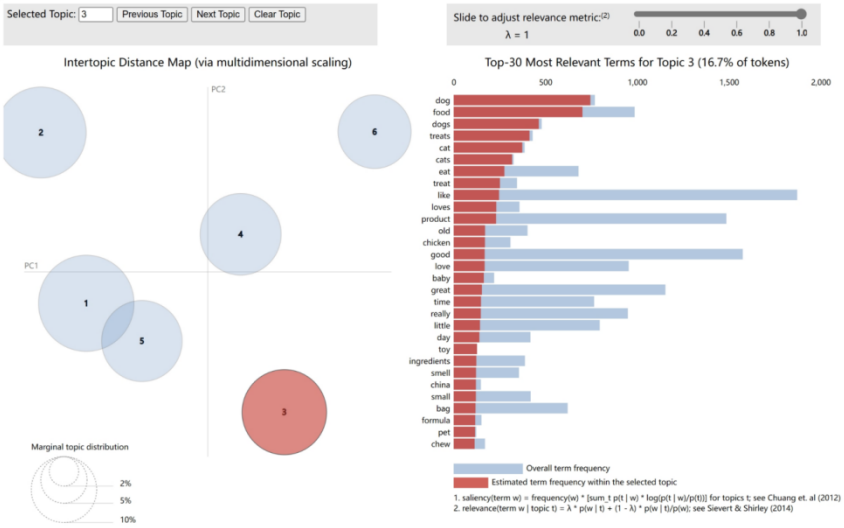

Figure 9. The LDA Model Themes Three: pet supplies.

4.2.4. Theme 4: Food Shopping Experience and Gift Purchasing. As shown in Figure 12. Theme 4 analysis focuses on the customer experience in purchasing food and gifts online, encompassing 15.3% of the dataset's vocabulary. Terms like 'like,' 'chips,' 'taste,' 'box,' and 'bag' highlight the importance of packaging and taste in customer purchase decisions. Frequently occurring terms in customer reviews such as 'good,' 'candy,' 'popcorn,' 'cookies,' and 'chocolate' reflect the preference and selection of various snack foods. Additionally, terms like 'buy,' 'bought,' 'ordered,' and 'sale' show the transactional process of online shopping. Terms like 'disappointed,' 'time,' 'bad,' 'tasted,' 'package,' and 'arrived' reveal customer expectations and actual experiences regarding product delivery time and quality, possibly relating to customer satisfaction and loyalty. 'Gift' and 'away' indicate that in choosing food as gifts, customers also consider their suitability and appeal as presents, suggesting that online retailers should consider offering more attractive packaging and gift services. In summary, LDA Theme 4 results indicate that in online food shopping and gift purchasing, product taste, packaging, delivery time, and suitability as gifts are key factors affecting customer purchase decisions. To enhance customer satisfaction and loyalty, retailers should focus on optimizing these aspects.

DOI: 10.54254/2755-2721/76/20240564

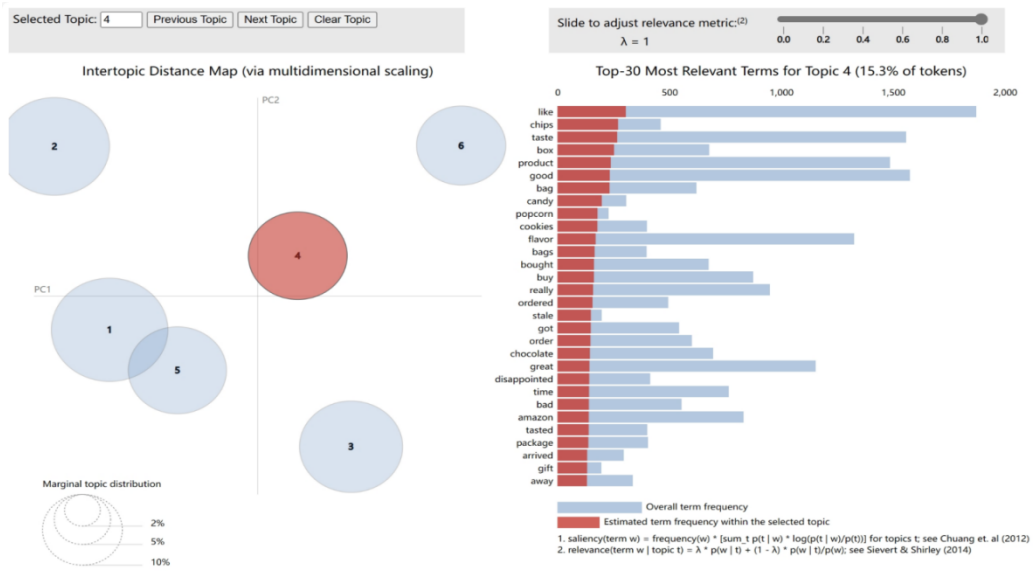

Figure 10. The LDA Model Themes Four: customer experience in purchasing food and gifts online.

4.2.5. Theme 5: Food Seasonings and Cooking Ingredients. As shown in Figure 13. In this study, applying machine learning to analyze Amazon product reviews, we revealed various factors influencing customer online shopping decisions. LDA Theme 5, encompassing 15.1% of the dataset's vocabulary, focuses on the selection of food seasonings and cooking ingredients. Terms in this theme such as 'sauce,' 'salt,' 'oil,' and 'spice' show customer preferences and emphasis on seasonings and cooking oils in food purchases. The choice of these ingredients reflects customer attention to food taste and cooking quality. Other terms like 'good,' 'flavor,' 'taste,' and 'great' demonstrate the significance of food taste in customer reviews. Terms like 'soup,' 'cheese,' 'rice,' and 'pasta' show diverse customer needs for staple and complementary foods. Additionally, terms like 'gluten,' 'milk,' and 'free' may point to customer attention to special dietary needs and food allergies. Terms like 'jerky,' 'noodles,' and 'chicken' show interest in specific food types. LDA Theme 5 analysis suggests that customers online not only focus on basic food taste and quality but also consider special dietary needs and food diversity. Thus, for retailers, highlighting these features in product descriptions and offering options for special diets can help meet diverse customer needs and influence their purchasing decisions.

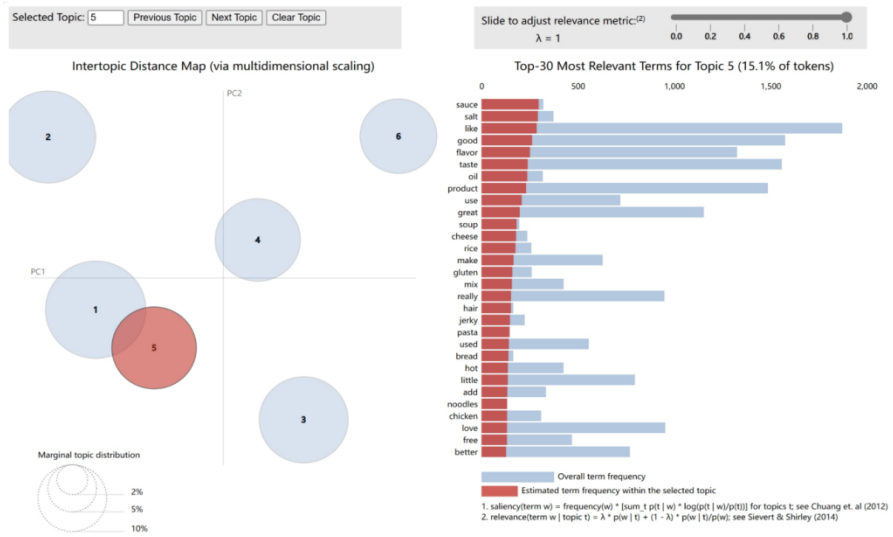

Figure 11. The LDA Model Themes Five: the selection of food seasonings and cooking ingredients.

DOI: 10.54254/2755-2721/76/20240564

4.2.6. Theme 6: Price Perception and Shopping Experience. As shown in Figure 14. In this machine learning analysis of Amazon product reviews, Theme 6 sheds light on factors considered by customers in evaluating their shopping experience, occupying 12.5% of the dataset's vocabulary. Theme 6 involves aspects such as pricing, purchasing process, and customer service. Significant terms in customer reviews like 'amazon,' 'price,' and 'deal' indicate that platform recognition and pricing are crucial in purchase decisions. These, along with terms like 'cheaper,' 'cost,' and 'discount,' emphasize customer sensitivity to price advantages. Moreover, terms such as 'order,' 'shipping,' 'received,' and 'purchase' reflect the importance customers place on the shopping process. Terms mentioned in reviews like 'service,' 'buying,' and 'love' highlight the role of customer service and brand loyalty in the shopping experience. Terms like 'store,' 'local,' 'grocery,' and 'cans' point to customer comparisons between online and traditional shopping, while terms such as 'pack,' 'box,' and 'time' may relate to product packaging and delivery efficiency. The analysis of Theme 6 indicates that in evaluating online shopping experiences, customers focus not only on price and discounts but also on convenience and quality of service. Therefore, retailers should focus on optimizing the shopping process, offering competitive pricing, and enhancing customer service experiences to improve online sales strategies.

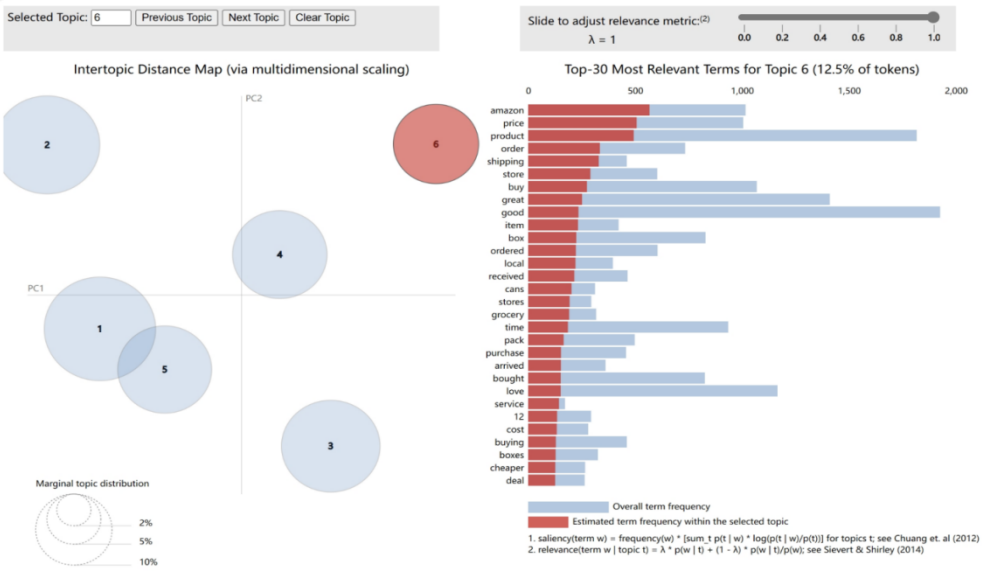

Figure 12. The LDA Model Themes Five: pricing, purchasing process, and customer service.

4.3. Word Cloud

The Word Cloud function refers to a specific tool used to discover a vector of length two to indicate the range of the word sizes and output a specific word classification as the form of a picture [17]. This research adopted the WordCloud function to map user comments to see the effect of sentiment analysis. The results are as follows: From the positive sentiment word cloud (Figure 5), the frequency of positive sentiment words such as 'good', 'cute', 'great' and so on is higher, which shows that sentiment analysis can better draw out the positive sentiment comments. Word cloud suggested that customers are much more likely to get attachments with the product outlooks. From the word cloud of negative sentiment comments (Figure 6), the frequency of negative sentiment words such as 'usual', 'bad', and 'normal' is higher, which shows that sentiment analysis can better extract the negative sentiment comments out of the negative sentiment comments. On the other hand, this phenomenon indicates that customers prefer products with certain functions or emotional connections [17].

DOI: 10.54254/2755-2721/76/20240564

Figure 13. Positive Words WordCloud.

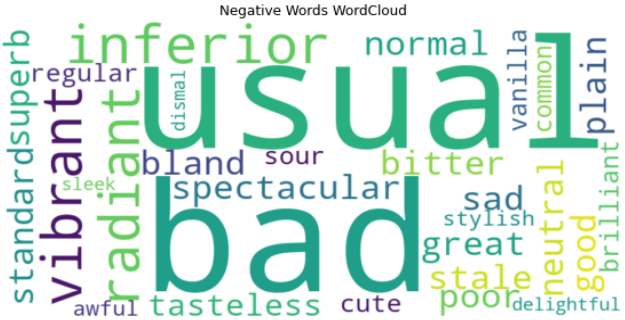

Figure 14. Negative Words WordCloud.

5. Conclusion

This work delved into the complexities of consumer behavior on Amazon, employing machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), and Latent Dirichlet Allocation (LDA). The outcomes of this research have shed light on the intricate landscape of sentiment analysis in e-commerce, particularly emphasizing the influence of product attributes like flavor, texture, and pricing on consumer choices. The performance of various machine learning algorithms in sentiment classification revealed nuanced insights. The Random Forest model, enhanced by the Tfidf Vectorizer, demonstrated notable precision and recall, effectively detecting diverse sentiment polarities. Conversely, Naive Bayes, though useful, showed limitations in precision and recall metrics. The choice of vectorization techniques, CountVectorizer or Tfidf Vectorizer, was found to significantly influence the performance, underlining the importance of method selection in sentiment analysis.

These results emphasize the significant impact of product attributes on consumer behavior. Sensory aspects such as flavor and texture, especially in food-related products, are critical in shaping customer preferences, highlighting the importance of sensory marketing. Similarly, sensitivity to pricing and discounts was evident, pointing to the necessity for competitive pricing strategies in e-commerce. This study contributes to the broader understanding of e-commerce strategy, showcasing the value of advanced analytics in decoding consumer behavior. Machine learning offers businesses insights to optimize their offerings and marketing strategies, potentially enhancing customer satisfaction and loyalty. However, it is acknowledged that this research has limitations, including the need for enhanced algorithmic precision and broader application across different e-commerce platforms. Future research should aim at expanding these methodologies, exploring more diverse data sets, and considering additional variables influencing online shopping behavior. In summary, this work represents a

DOI: 10.54254/2755-2721/76/20240564

foundational advancement in integrating analytics in e-commerce. It paves the way for future exploration in the field, highlighting the growing importance of machine learning in shaping the landscape of online shopping.

Acknowledgement
Ziqin Qin and Kefan Dong contributed equally to this work and should be considered co-first authors.

References

[1] Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon. in. Journal of retailing and Consumer Services, 45, 21-32.
[2] Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. International Journal of Information Management, 54, 102132.
[3] Singla, Z., Randhawa, S., & Jain, S. (2017). Sentiment analysis of customer product reviews using machine learning. Paper presented at the 2017 International Conference on Intelligent Computing and Control (I2C2), 1-5.
[4] Chang, V., Liu, L., Xu, Q., Li, T., & Hsu, C. (2023). An improved model for sentiment analysis on luxury hotel review. Expert Systems, 40(2), e12580.
[5] Chaffar, S. & Inkpen, D., Using a Heterogeneous Dataset for Emotion Analysis in Text, Advances in Artificial Intelligence, Canadian AI 2011, Lecture Notes in Computer Science, 6657, pp. 62-67, Springer, Berlin, Heidelberg, 2011.
[6] Y. Xiao, C. Qi and H. Leng, "Sentiment analysis of Amazon product reviews based on NLP," 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Changsha, China, 2021, pp. 1218-1221
[7] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE access, 8, 23522-23530.
[8] Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.
[9] Geetha, M. P., & Renuka, D. K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks, 2, 64-69.
[10] Gani, H., & Tomimatsu, K. (2019). Using Customer Emotional Experience from E-Commerce for Generating Natural Language Evaluation and Advice Reports on Game Products. Journal of ICT Research and Applications, 13(2), 145-161.
[11] Wickramasinghe, I., & Kalutarage, H. (2021). Naive bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293.
[12] Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology, 12(1), 1-7.
[13] Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.
[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
[15] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.
[16] Huang, L., Ma, J., & Chen, C. (2017, December). Topic detection from microblogs using T-LDA and perplexity. In 2017 24th Asia-Pacific software engineering conference workshops (APSECW) (pp. 71-77). IEEE.

DOI: 10.54254/2755-2721/76/20240564

[17] Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer Science Review, 41, 100413.