THE American Journal of Humanities and Social Sciences Research (THE AJHSSR) 2025
E-ISSN: 2581-8868
Volume-08, Issue-03, pp-199-204
www.theajhssr.com Crossref DOI: https://doi.org/10.56805/ajhssr
Research Paper Open Access

Investigating Sentiment Analysis Models for Online Reviews

[1]Dongwon Kim, [2]Yeonjoo Kim
[1]*(College of Liberal Arts/ Korea Maritime and Ocean University, Republic of Korea)*
[2]*(Department of English Language and Literature/ Korea Maritime and Ocean University, Republic of Korea)*

ABSTRACT

Online reviews are crucial for evaluating a company's value in the market and significantly impact its revenue. Therefore, sentiment analysis metrics from online reviews are vital indicators for predicting business success. This study focused on review texts from Hotels.com, a leading online hotel review platform. For polarity prediction of hotel reviews, this research employed four machine learning algorithms: Gradient Boosting Machine (GBM), and Light GBM, Logistic Regression (LR), Support Vector Machine (SVM). The performance evaluation results showed that Logistic Regression, SVM, and Light GBM algorithms achieved the highest accuracy at 0.87. This study contributes by quantifying unstructured text review data to predict ratings, enabling businesses, including startups, to effectively analyze customer feedback. Furthermore, it is expected to provide valuable insights that business operators can use to predict consumer behavior and formulate marketing strategies.

KEYWORDS – Forecasting algorithms, On-line reviews, Sentiment analysis, Social media.

1. INTRODUCTION

With the advancement of information and communication technologies, the digital economy mechanism has expanded, establishing the platform ecosystem as a core component of all industries and businesses. For instance, the platform economy, where economic activities between suppliers and demanders of goods and services occur around platforms, has rapidly grown in sectors like commerce and services. This is because companies selling products increasingly leverage e-commerce platforms as their primary sales channel, and customers similarly choose platforms as their main purchasing route [1]. As key stakeholders, including businesses and customers, actively participate in the platform ecosystem, platforms play an essential role in the modern business environment [2]. Notably, the COVID-19 pandemic, which solidified "non-face-to-face untact" as a sociocultural trend, further accelerated the expansion of the platform economy. As the digital economy expands and platforms become more important, the way we access and use information is also changing. Customers, in particular, can now easily get a wide variety of information through online platforms. This means that understanding the type and characteristics of information is crucial for corporate marketing strategies and consumer decision-making. This shift not only fuels the growth of platform-centric economic activities but also helps customers make more rational decisions by collecting both objective and subjective information about products and services. Consequently, research into the characteristics of information and its use in decision-making processes is becoming even more vital within the platform economy. Meanwhile, there are broadly two types of information: factual information, which presents objective characteristics of products and services (e.g., product functions), and subjective information, which conveys intangible attributes (e.g., feelings about a product) [3]. With the advancement of the internet, the development of various online platform-based media has created an environment where customers can more easily access subjective information, such as reviews from previous purchasers, in addition to factual information. Frederick et al. (2002) posited that in uncertain situations, customers undergo a logical judgment process based on both objective information (e.g., quantitative metrics or physical attributes) and subjective information like personally perceived utility to reduce risk and uncertainty [4]. They explained this decision-making through a Dual-Process Model, stating that subjective information, unlike impersonal sources, influences preferences and thus plays a role in purchase decisions. Raju et al. (1995) investigated the influence of objective knowledge, subjective knowledge, and usage experience on customer decision-making [5]. Their findings revealed that while these factors are interrelated, each element independently impacts the decision process.

Moreover, online reviews hold significant meaning because they contain diverse information, specifically evaluations of various elements related to user experience [6]. For example, a customer who prioritizes 'taste' in a restaurant review will likely comment on that aspect, whereas a customer valuing 'atmosphere' or 'service' will likely evaluate what matters to them, rather than the taste. This characteristic of subjective evaluation means that different customers prioritize different aspects, leading to the generation of a wide array of information that subsequent customers can reference. Online reviews are also a crucial source of information for businesses. Because they contain evaluations of products and services, they serve as vital data for exploring customer behavior and investigating market responses.

However, while online reviews effectively resolve information asymmetry, they also lead to a new problem: information overload, which increases the cost of searching for information [7]. Although customers benefit from the convenience of accessing diverse information through platforms, the growing volume of online reviews means it now takes time to find the specific information they want (e.g., factors like taste, price, or atmosphere in a restaurant).

From a business perspective, the increase in unstructured data like online reviews has resulted in significant costs and time expenditure during the extensive data analysis process. Hong et al. (2017) suggested that review length and ambiguity of content can affect review usefulness, while Mariani et al. (2023) explained that longer reviews tend to decrease in usefulness as information value [1][8].

Recently, attempts have been made to address these issues by leveraging machine learning to quantify online reviews into customer ratings, thereby enhancing their utility. Alslaity & Orji (2024) noted that quantified ratings make it easier to grasp product evaluation information. As this discussion gained prominence, research into whether quantified ratings accurately reflect the content of online reviews emerged as a significant issue [9]. Machine learning techniques offer a solution to this problem by proposing models that predict review text as ratings. Therefore, this study aims to propose a model that predicts online reviews as ratings using machine learning algorithms.

This study aims to perform Sentiment Analysis on review data from Hotels.com. The goal is to contribute to effectively analyzing customer feedback and providing predictable insights.

Specifically, we will explore models with high predictive performance and strive to improve model accuracy by utilizing six machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Machine (GBM), and LightGBM (Light Gradient Boosting Machine). Through this, we anticipate contributing to the development of sophisticated sentiment analysis models applicable across various industries that require customer feedback analysis.

2. THEORETICAL BACKGROUND

Prior research consistently highlights the pivotal role of online reviews in alleviating information asymmetry and enhancing utility.

From an information economics perspective, online reviews address the challenge of information asymmetry, where consumers struggle to obtain comprehensive information about product quality or alternatives. They effectively reduce the uncertainty stemming from a lack of information in purchasing decisions [10]. Customers often show a greater trust in reviews from other consumers than in information provided by suppliers. This manifests as a behavioral pattern where consumers actively seek useful insights from reviews when making purchasing decisions. According to "cue-summation theory," customers use these reviews as cues to expand their knowledge and make informed choices, making valuable reviews a crucial tool in the buying process.

Furthermore, reviews containing customer experiences exert a social influence, categorized as normative social influence, which strongly impacts potential buyers.

From a big data analytics methodology standpoint, studies underscore the importance of quantitatively assessing product strengths and weaknesses and service quality levels through review text analysis to predict customer purchase intent, satisfaction, and product preferences [11]. This has led to the development of various algorithms and methodologies in the field of sentiment analysis and machine learning. Examples include efforts to improve the accuracy of positive/negative classifications in movie reviews and the statistical and sentiment analysis of consumer product reviews, often filtering data into positive and negative categories.

These preceding studies demonstrate that online reviews significantly contribute to reducing uncertainty for customers during their decision-making process and providing businesses with valuable data for predicting customer behavior, enhancing the reliability of product and service evaluations, and developing effective marketing strategies and recommendation systems.

Recent research actively utilizes machine learning algorithms for sentiment analysis. Jemai et al. (2021) demonstrated that SVM and Random Forest are particularly effective for sentiment analysis, exhibiting high accuracy. While Naive Bayes offers fast computation, its performance can vary depending on data characteristics [12]. Veena et al. (2021) improved sentiment analysis accuracy to 0.84 using the VADER algorithm, which is highly regarded for its ability to precisely measure the emotional intensity and polarity of sentences [13]. Research using Yelp review data is also prevalent. Xu et al. (2015) compared Naive Bayes and Multiclass SVM with Yelp review data, concluding that Naive Bayes is more suitable for large-scale text classification, emphasizing the importance of algorithm selection based on data characteristics [14]. Similarly, Hemalatha et al. (2019) found that while Naive Bayes showed simple yet strong predictive power in sentiment analysis of Yelp review data, Linear SVC performed better in specific scenarios [15].

3. FRAMEWORK & MODELING

This study outlines the research procedure for modelling review sentiment analysis for digital platform complementors, as shown in Fig. 1.

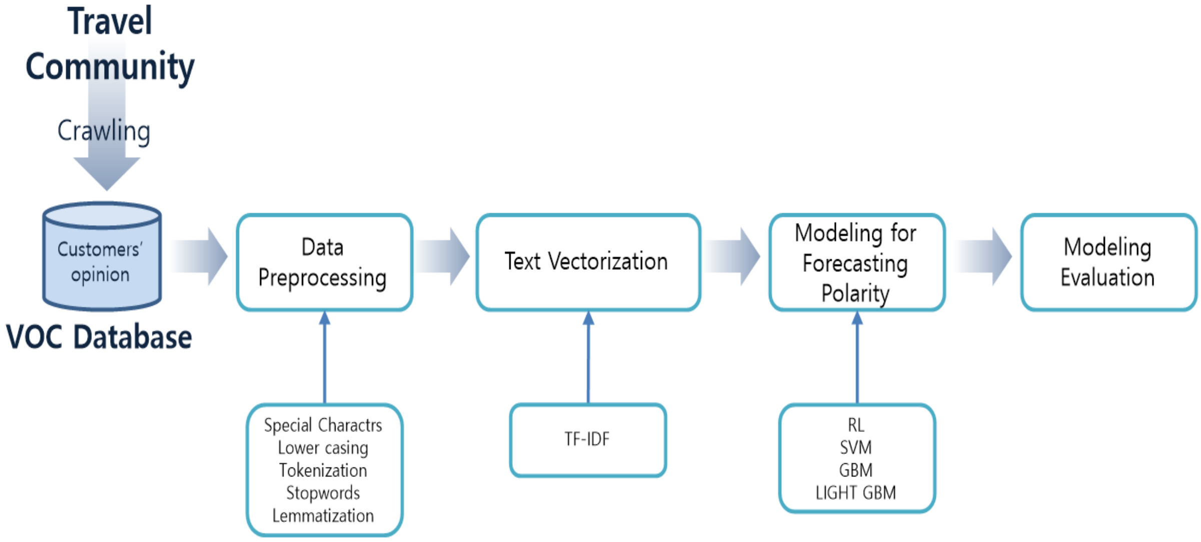

Fig. 1 Research Framework

4. DATA COLLECTION & PREPROCESSING

In this study, we employed a dataset from Hotels.com. To meet our research goals, we randomly selected 20,000 review records from over 600,000 available entries for analysis.

To balance the class distribution, we first performed sentiment classification and labeling of the review text. Review scores of 4 and 5 were categorized as the positive sentiment class, while 1 and 2 were assigned to the negative sentiment class. Reviews with a score of 3 were considered neutral and excluded from the analysis. An imbalanced class ratio is a critical consideration in many machine learning tasks, especially in classification problems. When one class is underrepresented compared to others, the model can become biased, struggling to predict the minority class and leading to poor performance for that specific group. To address this, we balanced the class ratios to ensure an even distribution among classes.

Since text data is difficult to directly apply to machine learning algorithms, a preprocessing stage to transform it into a suitable format is essential. In this study, we performed preprocessing on the review data using the SpaCy library. Specifically, the data was refined through the following steps:

First, to enhance model performance, we conducted Removing Special Characters. We then performed Lowercasing to prevent case confusion during text analysis, aiming to improve data consistency and analytical efficiency. The subsequent step involved Tokenization, which separates the text data into individual words or sentences. Additionally, we performed Stop Words Removal to eliminate insignificant vocabulary, such as articles and prepositions, that are meaningless for analysis. We also conducted Lemmatization, the process of converting words to their base or root form.

To leverage the text data for machine learning algorithms, we transformed it into numerical data using TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a method that calculates the weight of each word by simultaneously reflecting its frequency and its importance within the document. Through this approach, we extracted important keywords from the text and converted them into a vector form to be used as input for the machine learning model.

5. MODELING FOR FORECASTING POLARITY

In this study, we modeled the prediction of online review ratings using a total of four machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Machine (GBM), and LightGBM.

SVM is a supervised learning algorithm used for classifying given data into specific classes. It works by mapping data into a high-dimensional space to find the optimal decision boundary that maximizes the margin between classes.

Logistic Regression is an algorithm designed to solve classification problems. While based on linear regression, it passes its results through a logistic function (Sigmoid function) to yield values between 0 and 1. Consequently, it's primarily used for binary classification problems and provides probabilistic predictions.

GBM (Gradient Boosting Machine) operates by sequentially training weak learners, adding new learners in a way that aims to reduce the errors of the preceding ones. It is primarily based on decision trees. At each step, it corrects errors using the gradients of the loss function, thereby enhancing performance.

LightGBM is also based on the gradient boosting framework but is faster and uses less memory. Its tree growth method is leaf-wise rather than depth-based, making it more effective for larger datasets. However, with smaller datasets, there is an increased risk of overfitting.

simultaneously reflecting its frequency and its importance within the document. Through this approach, we extracted important keywords from the text and converted them into a vector form to be used as input for the machine learning model.

6. MODELING EVALUATION

To evaluate the performance of our experimental models, this study utilized a Confusion Matrix. The Confusion Matrix is a widely used tool for assessing the performance of classification models, allowing for a visual comparison of actual values against predicted values. This information consists of the following four elements: True Positive (TP): The model correctly predicted positive data as positive.

- True Negative (TN): The model correctly predicted negative data as negative.
- False Positive (FP): The model incorrectly predicted negative data as positive (Type I Error).
- False Negative (FN): The model incorrectly predicted positive data as negative (Type II Error).

Based on these four elements of the confusion matrix, performance metrics for classification models can be calculated. In this study, we utilized Accuracy, Precision, Recall, and F1 score. The calculation methods for these metrics are described in the following Table 1.

Table 1. Model Performance Evaluation Metrics

Metric	Formula
Accuracy	TP+TN / (TP+FP+TN+FN)
Precision	TP / (TP+FP)
Recall	TP / (TP+FN)
F1- Score	2* (Precision*Recall)/(Precision+Recall)

7. ANALYSIS RESULTS

The sentiment analysis performance evaluation experiments in this study were conducted to compare the predictive performance of various machine learning models based on review data, and to select the optimal model through hyperparameter tuning and 3-Fold Cross Validation. Cross-validation was used to enhance the reliability of model performance, a method that divides the dataset into three subsets (folds) and repeatedly performs training and evaluation on each fold. This process allows for the verification of the model's generalization performance. Furthermore, Accuracy, Precision, Recall, and F1 Score values were calculated using a weighted average to ensure that performance evaluation was accurately reflected even with imbalanced data. The analysis results are presented in Table 2 below.

Table 2. 3-Fold Cross Validation

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.87	0.87	0.87	0.87
SVM	0.87	0.88	0.87	0.88
GBM	0.81	0.81	0.81	0.80
LightGBM	0.87	0.87	0.86	0.87

LR, SVM, and LightGBM all demonstrated the highest performance in this study, each achieving an Accuracy of 0.87. Notably, all three algorithms also showed consistent performance across Precision, Recall, and F1 Score, ranging between 0.86 and 0.88. This indicates a well-balanced predictive capability for both positive and negative sentiment classes. Regarding execution speed, Logistic Regression and SVM completed their runs faster than LightGBM. This suggests that from the perspective of computing resource utilization, employing Logistic Regression or SVM could be more advantageous. GBM achieved an Accuracy of 0.81, with its Precision, Recall, and F1-score metrics also performing within the 0.80-0.81 range. This outcome suggests that GBM faced some challenges in maintaining consistency in predictions for both positive and negative classes.

8. CONCLUSION

This study proposed a method for quantitatively evaluating and predicting customer feedback through sentiment analysis modeling using online review data. We built polarity prediction models using machine learning algorithms such as Logistic Regression, SVM, and LightGBM. The results showed that Logistic Regression and SVM algorithms achieved high accuracy. Furthermore, model performance was enhanced through hyperparameter tuning and cross-validation. While the models demonstrated high accuracy in binary classification, we identified a limitation where they showed relatively lower accuracy in multiclass classification. These findings contribute to comparing the characteristics and performance of various sentiment analysis models and understanding the strengths and weaknesses of each algorithm.

By analyzing online reviews, companies can quickly grasp customer needs and market changes. This allows them to enhance customer satisfaction by encouraging positive reviews and effectively managing negative ones. The results of this study can be usefully applied to provide a model that helps businesses more systematically understand customer feedback through online reviews and formulate marketing strategies based on this understanding. Notably, the sentiment analysis model proposed in this study is significant as it can be utilized for reputation analysis and the development of personalized recommendation systems across various fields, including movies, books, and music.

REFERENCES

1. Mariani, M. M., Borghi, M., & Laker, B.(2023). Do submission devices influence online review ratings differently across different types of platforms? A big data analysis. Technological Forecasting and Social Change, 189, 1-12.
2. Choi, H. S., & Leon, S.(2020). An empirical investigation of online review helpfulness: A big data perspective. Decision Support Systems, 139, 1-12.
3. Deutsch, M., & Gerard, H. B.(1955). A study of normative and informational social influences upon individual judgment. The Journal of Abnormal and Social Psychology, 51(3), 629-636.
4. Frederick, S., Loewenstein, G., & O'donoghue, T.(2002). Time discounting and time preference: A critical review. Journal of Economic Literature, 40(2), 351-401.
5. Raju, P. S., Lonial, S. C., & Mangold, W. G.(1995). Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: An exploratory investigation. Journal of Consumer Psychology, 4(2), 153-180.
6. Thakur, R.(2018). Customer engagement and online reviews. Journal of Retailing and Consumer Services, 41, 48-59. Tripathy, A., & Rath, S. K.(2017). Classification of sentiment of reviews using supervised machine learning techniques. International Journal of Rough Sets and Data Analysis, 4(1), 56-74.
7. Wang, F., Du, Z., & Wang, S.(2023). Information multidimensionality in online customer reviews. Journal of Business Research, 159, 1-15.
8. Hong, H., Xu, D., Wang, G. A., & Fan, W.(2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. Decision Support Systems, 102, 1-11.
9. Alslaity, A., & Orji, R.(2024). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour & Information Technology, 43(1), 139-164.
10. Cialdini, R. B., & Goldstein, N.(2009). Normative influences on consumption and conservation behaviors. Social Psychology of Consumer Behavior, 273-296.
11. Hemalatha, S., & Ramathmika, R.(2019). Sentiment analysis of yelp reviews by machine learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 700-704). IEEE.
12. Jemai, F., Hayouni, M., & Baccar, S. (2021, June). Sentiment analysis using machine learning algorithms. In 2021 International Wireless Communications and Mobile Computing (IWCMC) (pp. 775-779). IEEE.
13. Veena, G., Vinayak, A., & Nair, A. J.(2021). Sentiment Analysis using Improved Vader and Dependency Parsing. 2021 2nd Global Conference for Advancement in Technology (GCAT). IEEE.
14. Xu, Y., Wu, X., & Wang, Q.(2015). Sentiment analysis of yelp's ratings based on text reviews. 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 17(1).
15. Hemalatha, S., & Ramathmika, R.(2019). Sentiment analysis of yelp reviews by machine learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 700-704). IEEE.