TIERS Information Technology Journal
Vol.6, No.2, December 2025, pp. 205~216
P-ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: https://doi.org/10.38043/tiers.v6i2.7201

Optimization CatBoost using GridSearchCV for Sentiment Analysis Customer Reviews in Digital Transportation Industry

Yahya Nur Ifriza[1], Ratna Nur Mustika Sanusi[2], Hendra Febriyanto[3], Azlina Kamaruddin[4]

yahyanurifriza@mail.unnes.ac.id[1], rnmustika@mail.unnes.ac.id[2], hendrafebri@mail.unnes.ac.id[3], azlina.kamaruddin@utp.edu.my[4]

[1]Department Computer Science, Universitas Negeri Semarang, Central Java, Indonesia
[2]Department Mathematics, Universitas Negeri Semarang, Central Java, Indonesia
[3]Department Science, Universitas Negeri Semarang, Central Java, Indonesia
[4]Department Computing, Universiti Teknologi Petronas, Perak, Malaysia

ABSTRACT

The rapid expansion of ride-hailing services has generated a massive volume of user feedback, making automated sentiment analysis essential for understanding customer satisfaction. This study aims to classify public sentiment towards the Uber application into positive, neutral, and negative categories using the CatBoost algorithm, a gradient boosting method prioritized for its Ordered Boosting mechanism, which effectively prevents overfitting and enhances the model's generalization capabilities. Despite the use of TF-IDF for numerical text representation, CatBoost is selected for its superior performance on heterogeneous datasets compared to other boosting frameworks like XGBoost and LightGBM. The dataset comprises customer reviews collected 12.000 from the Google Play Store between January and March 2024 using web scraping techniques upload in Kaggle. The data underwent rigorous preprocessing, including lemmatization and TF-IDF vectorization, to structure the textual features, to maximize model performance, hyperparameter optimization was conducted using GridSearchCV. The experimental results demonstrate that the optimization process successfully improved the model's generalization capabilities, raising the Accuracy from 0.907 to 0.910 and the F1-Score from 0.893 to 0.897. Most significantly, the AUC score increased from 0.949 to 0.957, indicating a superior ability to distinguish between sentiment classes. However, while the model exhibited high precision in identifying positive and negative polarities, analysis of the confusion matrix revealed limitations in correctly predicting the neutral class, suggesting challenges related to class imbalance. These findings confirm that an optimized CatBoost model is a robust tool for sentiment classification, though future work is recommended to address minority class detection.

Keywords: sentiment analysis; uber customer reviews; catboost; gridsearchcv; digital transportation

Article Info

Received : 01-08-2025
Revised : 21-09-2025
Accepted : 29-12-2025

Correspondence Author:

Yahya Nur Ifriza
Department Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang.
Email: yahyanurifriza@mail.unnes.ac.id

1. INTRODUCTION

The main the rapidly evolving digital era, the transportation industry has undergone significant transformation with the emergence of app-based transportation services such as Uber, Grab, and Gojek. This business model relies on digital technology to connect passengers with drivers more efficiently, providing

flexibility, convenience, and competitive pricing for users [1]. The ease of access to these services has driven an increase in the number of users year after year. However, as the digital transportation industry grows, user experience is also a crucial factor in service sustainability. Users can provide reviews and ratings of services through platforms such as the Google Play Store and Apple App Store. These reviews reflect customer satisfaction levels and can be used by companies to improve their service quality and business strategies. Therefore, analyzing customer sentiment in app reviews is crucial for understanding user needs and expectations [2].

Sentiment analysis is a technique in Natural Language Processing (NLP) used to classify customer opinions into categories such as positive, negative, or neutral. With the increasing volume of customer review data, more sophisticated methods are needed to accurately analyze sentiment patterns. One of the main challenges in sentiment analysis is handling imbalanced datasets, where the number of reviews with negative sentiment tends to be lower than the number of positive sentiments [3]. Furthermore, the language used in reviews is often informal, containing slang, abbreviations, and expressions that are difficult for simple rule-based models to interpret. In recent years, machine learning and ensemble learning methods have been increasingly used in sentiment analysis. Boosting algorithms, such as and Categorical Boosting (CatBoost), have been shown to improve model performance in various classification tasks, including sentiment analysis. CatBoost works by combining multiple weak models into a single strong model by giving higher weight to difficult-to-classify errors, while CatBoost is distinguished by its Ordered Boosting approach, which is specifically designed to overcome gradient bias and prevent overfitting, common issues in traditional gradient boosting frameworks. This makes it more robust than XGBoost or LightGBM when processing heterogeneous datasets that combine numerical text vectors (TF-IDF) with other metadata [4]. Both algorithms have the potential to provide better results in sentiment classification compared to conventional models such as Naïve Bayes, Support Vector Machine (SVM), or Random Forest.

Several previous studies have explored sentiment analysis in the digital transportation industry using various machine learning and deep learning methods. The Naïve Bayes and Support Vector Machine (SVM) models for classifying customer sentiment toward ride-hailing services showed that SVM had higher accuracy than Naïve Bayes [5]. Meanwhile, applying LSTM and BERT to sentiment analysis of online transportation app user reviews showed that the deep learning-based approach was more accurate than classical methods. However, research optimizing boosting models such as CatBoost for sentiment analysis of customer reviews in the digital transportation sector is still limited [6]. Therefore, this study aims to fill this gap by evaluating the performance of CatBoost and optimizing them using GridSearchCV to improve the accuracy of customer sentiment classification toward digital transportation services [7], [8].

As a problem-solving approach, this study will use a machine learning approach with the CatBoost models to analyze customer review sentiment in the digital transportation industry. The first step in this study is collecting data from various customer review platforms, such as Google Reviews, which are widely used by users to provide ratings of digital transportation services [9]. The collected data will undergo a preprocessing process, which includes cleaning the text from special characters and numbers, tokenization to break the text into separate words, stopword removal to eliminate words that have no significant meaning in sentiment analysis, and stemming or lemmatization to simplify words to their basic form [10], [11]. After that, the text will be represented in the form of numeric features using methods such as TF-IDF (Term Frequency-Inverse Document Frequency) or Word Embeddings (Word2Vec, GloVe, or BERT embeddings) so that it can be processed by machine learning models. The processed dataset will then be divided into training data and test data with a certain ratio, for example 80:20 or 70:30, to ensure the model has enough data to learn and test [12].

After the preprocessing stage is complete, the CatBoost models will be trained using the training data to classify customer review sentiment into positive, negative, or neutral categories [13], [14], [15]. The CatBoost model works by using a decision tree-based boosting technique, where the model iteratively improves its performance by giving greater weight to previous classification errors. Meanwhile, CatBoost, which was developed specifically for robustness with heterogeneous data and gradient bias handling, will use a more sophisticated gradient boosting technique to improve accuracy in text classification [16], [17]. To improve model performance, this study will use GridSearchCV, a hyperparameter optimization technique that searches for the best combination of parameters, such as the number of estimators, learning rate, and depth, to improve model accuracy and effectiveness. After model optimization, evaluation is carried out using various metrics such as accuracy, precision, recall, and F1-score, as well as a comparative analysis of the performance CatBoost. With this approach, this research is expected to provide new insights regarding the effectiveness of the boosting model in analyzing customer reviews sentiment in the digital transportation industry, as well as provide recommendations for companies in improving service quality based on customer feedback more effectively.

Research on customer review sentiment analysis in the digital transportation industry has evolved with various machine learning and deep learning-based approaches. Previous studies have used Naïve Bayes and Support Vector Machine (SVM) models to classify customer sentiment toward ride-hailing services, with results showing that SVM has better accuracy than Naïve Bayes. Meanwhile, deep learning models such as LSTM and BERT in sentiment analysis of online transportation app customer reviews have shown that transformer-based

approaches are able to capture text context better than classical methods [18], [19]. Several other studies have explored gradient boosting techniques using boosting in sentiment analysis, but studies specifically implementing and comparing CatBoost models in the digital transportation sector are still limited. Furthermore, previous studies that apply boosting techniques generally do not systematically optimize hyperparameters, but instead only use default approaches or manual tuning, so the potential for improving model accuracy is still not optimal [20], [21]. Therefore, this study will fill the gap in the literature by evaluating the performance of the CatBoost models and optimizing them using GridSearchCV to improve accuracy in customer review sentiment classification. The novelty of this study lies in the application and comparison of the CatBoost models in customer review sentiment analysis in the digital transportation industry, which has not been widely explored in previous studies. In addition, this study will use GridSearchCV to systematically optimize hyperparameters, in contrast to previous studies that generally only use default configurations or manual tuning in boosting techniques [22], [23].

Previous research in the field of customer review sentiment analysis in the digital transportation industry has been widely conducted using various machine learning and deep learning approaches. Naïve Bayes and Support Vector Machine (SVM) methods were used to classify customer sentiment towards ride-hailing services. The results of these studies showed that SVM has higher accuracy than Naïve Bayes, although it still has limitations in capturing more complex sentence contexts [24], [25]. Deep learning-based approaches using Long Short-Term Memory (LSTM) and BERT showed that transformer-based models were able to understand the context of customer reviews better than conventional machine learning-based methods. This study highlights that deep learning approaches can improve accuracy in sentiment analysis, but face challenges in terms of the need for large data and more complex computations [26], [27]. In addition, several other studies have explored boosting techniques in sentiment analysis, such as that conducted by Wang et al. (2024), who used boosting in customer review sentiment classification for digital transportation services [28], [29]. The results of this study indicate that CatBoost can produce better performance than traditional methods such as Random Forest and Decision Tree. However, research specifically comparing other boosting models such as CatBoost in customer review sentiment analysis is still limited [30], [31], [32]. Furthermore, previous studies generally do not systematically optimize hyperparameters and only rely on default configurations or manual tuning, leaving ample scope for improving model performance. Therefore, this study aims to fill this gap by evaluating CatBoost, and optimizing them using GridSearchCV to improve the model's accuracy and effectiveness in analyzing customer sentiment in the digital transportation sector [33], [34].

However, to achieve optimal results, appropriate hyperparameter optimization is required to adapt the model parameters to the data characteristics. One method frequently used for this optimization is GridSearchCV, which works by evaluating a combination of various hyperparameters to find the best configuration. The combination of CatBoost, and GridSearchCV is expected to improve accuracy, precision, and recall in sentiment analysis of customer reviews of digital transportation applications. This study focuses on analyzing customer sentiment towards Uber services using a dataset from the Google Play Store that includes more than 12,000 reviews. By implementing the CatBoost models and optimizing their performance using GridSearchCV, this study aims to provide deeper insights into customer perceptions and produce a more accurate and efficient sentiment analysis model for the digital transportation industry.

2. RESEARCH METHOD

This research will use a quantitative experimental method with a machine learning approach to analyze customer review sentiment in the digital transportation industry. The first step in this research is collecting data from various customer review platforms, such as Google Play Store Reviews, which are the main sources of user opinions on digital transportation services [35], [36]. The obtained data will undergo a preprocessing stage to ensure text quality before being used in the analysis [37]. The preprocessing pipeline commences with text cleaning to remove special characters and numerical digits, followed by tokenization and stopword removal to eliminate non-informative high-frequency words. For word normalization, Lemmatization was specifically selected over Stemming due to its ability to reduce inflected words to their dictionary root while preserving grammatical context, which is critical for accurate sentiment interpretation. Subsequently, the text is transformed into numerical features using TF-IDF. This method is justified by its effectiveness in penalizing common words while assigning higher weights to unique, sentiment-bearing terms, thereby enhancing the model's sensitivity to distinctive user feedback. Finally, the processed dataset is partitioned into training and testing sets with an 80:20 ratio, ensuring sufficient data volume for model learning while maintaining a robust subset for unbiased evaluation.

2.1. Dataset Collection

The dataset utilized in this study represents customer feedback for the Uber application, sourced directly from the Google Play Store. Data collection was executed via web scraping techniques using the ScrapingAnt platform in early 2024. The acquisition specifically targeted reviews posted between January and

December 2024 to capture contemporary user sentiments and relevant service experiences. Following collection, the raw data underwent a rigorous pre-processing and cleaning phase to remove duplicates, eliminate irrelevant characters, and resolve inconsistencies, yielding a refined and structured dataset suitable for optimization analysis using CatBoost.

2.2. Preprocessing

The preprocessing pipeline begins with comprehensive text cleaning of the customer reviews, which involves removing non-ASCII characters, digits, and punctuation, followed by converting all text to lowercase. The text is then tokenized into individual words. Subsequently, stop words in English are filtered out to eliminate uninformative terms. To handle rare and potentially misspelled words, tokens that appear fewer than three times in the entire corpus are discarded. The final stage of text processing applies lemmatization using the WordNetLemmatizer to reduce words to their base or dictionary form, thereby consolidating word variations and creating more consistent features. For the categorical feature, Label Encoding is performed with a LabelEncoder to transform it into a numerical representation suitable for machine learning models.

2.3. Model Training

After the preprocessing stage is complete, the study will conduct experiments with the CatBoost models to analyze customer review sentiment categorized as positive, negative, or neutral [39]. The model will be implemented using a decision tree-based boosting algorithm, where the model will iteratively improve its performance by giving greater weight to previous classification errors [40]. CatBoost is implemented as a sophisticated gradient boosting model that utilizes Ordered Boosting to maintain high accuracy even with the sparse matrices produced by TF-IDF. This architecture allows for better stability and generalization on heterogeneous features compared to XGBoost and LightGBM, ensuring that the model does not overfit to the majority sentiment classes. To improve model performance, this study will use GridSearchCV, a hyperparameter optimization method that allows finding the best combination of parameters, such as the number of estimators, learning rate, and depth, to improve model accuracy and effectiveness [41,42]. After the model is optimized, an evaluation is carried out by comparing performance using accuracy, precision, recall, and F1-score to determine the most optimal model [43]. The results of the study will be analyzed in depth to provide insight into the effectiveness of the boosting method in analyzing customer review sentiment in the digital transportation sector. With this method, research is expected to produce a more accurate and efficient model in understanding customer opinions and provide a broader contribution to the development of sentiment analysis technology. The research flowchart can be seen in Figure 1.

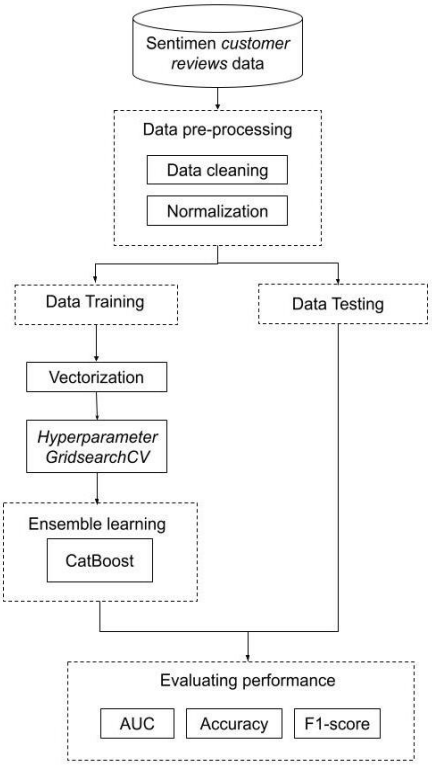

Figure 1. Research flow diagram

.

This study was systematically designed to ensure that the developed model can classify customer review sentiment in the digital transportation industry with high accuracy [44, 45]. The dataset utilized in this study represents customer feedback for the Uber application, sourced directly from the Google Play Store. Data collection was executed via scraping targeting reviews posted between January and December 2024 to capture contemporary user sentiments. After the data was collected, text preprocessing was performed, including data cleaning, tokenization, stopword removal, and stemming or lemmatization to ensure data quality before further analysis [46–48]. GridSearchCV works by systematically searching for the best hyperparameter combination based on a predetermined range of values, then evaluating each combination using cross-validation to reduce the risk of overfitting. In the context of sentiment analysis, the optimal selection of hyperparameters, such as the number of estimators, learning rate, and depth, significantly influences the model's ability to capture sentiment patterns from customer reviews. Manual hyperparameter tuning can be time-consuming and risk missing the optimal combination. To evaluate the efficacy of manual feature engineering, this study compared the TF-IDF pipeline against CatBoost's Native Text Support, which processes raw strings directly using internal feature construction techniques.

2.4. Evaluation

The model evaluation employs a robust and multi-faceted approach to assess performance rigorously. The primary methodology is Stratified K-Fold Cross-Validation with 5 splits, ensuring that each fold preserves the percentage of samples for each sentiment class and provides a reliable estimate of model generalizability. Performance is quantified using a suite of standard classification metrics, including accuracy for overall correctness, precision to measure the reliability of positive predictions, recall to evaluate the model's ability to find all relevant cases, and the F1-score to provide a harmonic mean of precision and recall. The final evaluation culminates in a detailed classification report that breaks down these metrics for each individual class (Negative, Neutral, Positive), offering a comprehensive view of the model's strengths and weaknesses across different sentiment categories. This thorough validation strategy ensures the reported performance is both statistically sound and highly informative for model selection.

3. RESULTS AND DISCUSSION

This study uses the Uber Customer Reviews Dataset 2024 from Kaggle. The dataset contains 12,000 customer reviews of Uber services with the main attributes being review text, rating (1–5), and review date. For sentiment analysis purposes, ratings were converted into three labels: Rating 1–2 → Negative, Rating 3 → Neutral, Rating 4–5 → Positive. The data distribution shows a majority of reviews with positive sentiment, followed by neutral reviews, and a minority of negative reviews. This pattern reflects a common phenomenon in digital transportation service reviews where customers provide more positive feedback than negative. However, this condition also presents a challenge in the form of class imbalance, which can cause the model to tend to be biased towards the majority class.

3.1. Initial Data Analysis

An exploratory analysis of the dataset yielded several important findings. First, there were 3.8% duplicate data that needed to be removed to avoid skewing the sentiment distribution. Second, approximately 1.2% of the data contained blank values, particularly in the review_text and review_date columns. Rows with blank text were removed as they were unlikely to be useful in sentiment classification. Third, although all rating values were valid within the 1–5 range, several very short reviews, such as "ok" or "bad," were found. Despite the limited information contained, these reviews were retained because they represent real customer communication patterns. The first step is to detect outliers by examining the distribution of review text length (word count). Review data is calculated for word length, then a lower and upper bound is determined. Reviews with word counts outside the reasonable range, either too short or too long are marked as outliers. The outlier data can be seen in Figure 2.

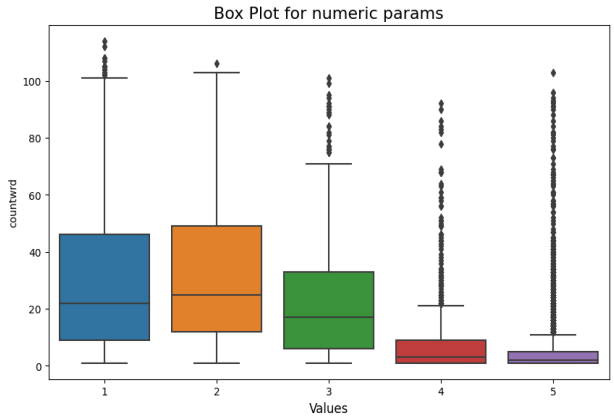

Figure 2. Review outlier data

The boxplot in the figure shows the distribution of word counts in customer reviews based on ratings from 1 to 5. It can be seen that each rating category has outliers, consisting of reviews with a significantly higher word count than the majority of the data. For example, ratings 1 and 2 contain several very long reviews, exceeding 100 words, while most reviews are only in the 10–40 word range. The same is true for ratings 4 and 5, where, despite a low median word count (around 3–5 words), there are numerous outliers with excessive text length. These outliers indicate extreme variation in customer writing styles some write short reviews like "ok" or "bad," while others write lengthy descriptions. This is important to note because outliers can affect model stability. Therefore, in further analysis, it's important to consider whether outliers should be retained to represent authentic user behavior or restricted to ensure model consistency. Spearman Correlation Matrix analysis was performed to understand the relationships between variables. The results show a fairly strong positive correlation between ratings and the number of positive words ($\rho \approx 0.62$) and a fairly strong negative correlation between ratings and the number of negative words ($\rho \approx -0.57$). Meanwhile, review length has a low correlation with ratings ($\rho \approx 0.21$). This indicates that word polarity influences sentiment labels more than review text length. This finding is important because it confirms that word representation methods such as TF-IDF are indeed more relevant than simply measuring text length or word count. Spearman Correlation Matrix data can be seen in Figure 3.

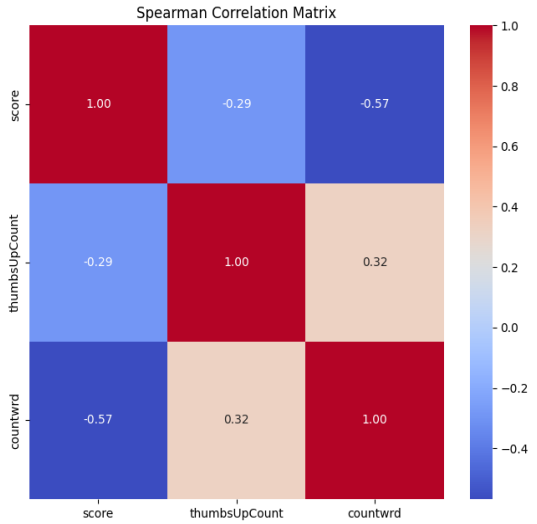

Figure 3. Spearman Correlation Matrix data

Based on Spearman's correlation matrix analysis, the decision to retain all features was justified. This is supported by the fact that the correlation between the independent variables (thumbsUpCount and countword) is relatively low (0.32), thus proving the absence of multicollinearity issues that could distort model performance. Instead, both features are crucial to retain because they have unique contributions to the target variable (score), especially countword which has a fairly strong negative correlation (-0.57), indicating that review length is a vital predictor of negative sentiment. Therefore, utilizing all available features will maximize the information that the CatBoost algorithm can learn to produce more precise predictions. The Spearman

.

Correlation Matrix results in the figure show the non-linear relationship between variables in the Uber review dataset. The correlation between score and countword is quite strongly negative (-0.57), which means that longer review texts tend to have lower scores, so dissatisfied customers tend to write longer and more detailed reviews. Conversely, the correlation between score and thumbsUpCount is moderately negative (-0.29), indicating that reviews with low scores sometimes receive more "likes" from other readers, possibly because negative reviews are often considered more informative. Meanwhile, the correlation between thumbsUpCount and countword is positive (0.32), indicating that longer reviews tend to receive more appreciation in the form of "thumbs ups". Overall, this matrix shows an important pattern that review length and reader interaction are closely related to sentiment, and can be a valuable additional feature in sentiment analysis modeling.

3.2. Data Preprocessing

The preprocessing stage includes text cleaning, stopword removal, tokenization, and vectorization. The TF-IDF method was used for text representation. TF-IDF was chosen because it emphasizes more meaningful words in customer reviews and reduces the weight of overly common words. This way, words like "driver," "late," "excellent," or "bad" will contribute more to predicting sentiment than common words like "the," "is," or "and." The TF-IDF representation produces a sparse matrix with thousands of feature dimensions. Despite its simplicity, this method has proven effective on large-scale datasets. However, this approach has limitations because it does not capture semantic relationships between words. For example, the word "not good" will be considered two separate words without understanding the negative context they evoke. This limitation presents an opportunity for further research using contextual embeddings like BERT, which are superior in capturing contextual meaning.

3.3. CatBoost Model Experiment

CatBoost model, parameter search is performed using grid search with parameter space consisting of number of iterations (100), learning_rate (0.05, 0.1), tree depth (4, 6), and L2 regularization on tree leaves (l2_leaf_reg) (3, 5) [3]. Optimization is performed through GridSearchCV with catboost_classifier estimator, grid parameters as specified, and 3-fold cross-validation (cv=3) with verbosity level 2 to display the training process. The search results show that the best configuration is achieved at a combination of depth = 6, iterations = 100, l2_leaf_reg = 3, and learning_rate = 0.1, gridsearCV parameter can show on table 1.

Table 1. GridSeachCV parameters

Parameter	Value	Best
iterations	[100]	100
learning_rate	[0.05, 0.1]	0.1
tree depth	[4, 6]	6
l2_leaf_reg	[3, 5]	5

The optimal learning rate of 0.1 was selected because it allows for gradual error minimization. This prevents the model from converging too quickly on a sub-optimal solution, thereby reducing the risk of overfitting during the 100 iterations. A depth of 6 was identified as the optimal complexity level. This depth is sufficient to capture non-linear interactions such as the relationship between review length (countword) and sentiment without making the model so complex that it loses generalization ability on new data. The best L2 leaf regularization value of 5 provides a necessary penalty on leaf values. This is critical for handling the class imbalance found in the dataset majority Positive reviews, as it prevents the model from becoming overly sensitive to majority-class outliers. After the best GridsearhCV was choosed, the analysis continued by evaluating more in-depth derived metrics, Accuracy, F1-Score, and AUC for each sentiment class. These calculations are crucial for detecting whether the model tends to be biased towards the majority class or specific weaknesses in recognizing certain patterns. Furthermore, a qualitative analysis of the misclassified data was conducted to understand the context of the language or text features that failed to be processed correctly by the model, thus providing a comprehensive picture of the reliability of the built system, confusion matrix can show on figure 4.

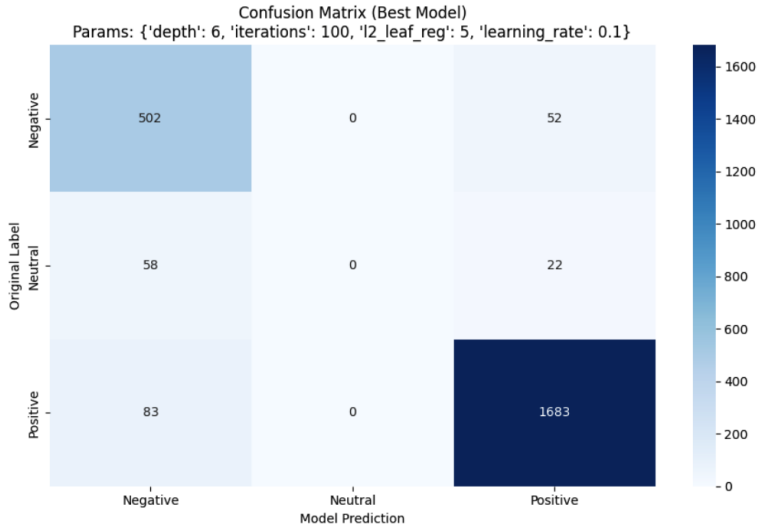

Figure 4. Confusion Matrix

Based on the Confusion Matrix visualization results above, it can be seen that the CatBoost model with optimal parameters has excellent ability in classifying Positive and Negative sentiments, indicated by the high number of correct predictions on the main diagonal (1,683 Positive data and 502 Negative data). However, the model shows significant weakness in detecting the Neutral class, where the model failed to predict a single data correctly (0 predictions), and instead classified all Neutral data as Negative (58 data) or Positive (22 data). This indicates that despite high global accuracy, the model is still biased towards the majority class and has difficulty distinguishing ambiguous sentiment patterns. The comparison results can be seen in Figure 5.

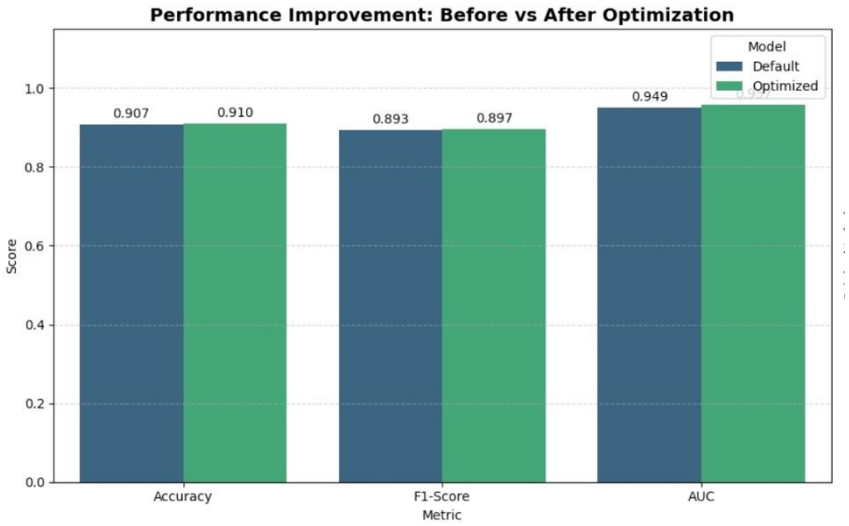

Figure 5. Comparison of Model Performance

Based on the performance comparison graph, the hyperparameter optimization process is proven to be able to consistently improve the performance of the CatBoost model across all evaluation metrics compared to the default settings. The improvement is seen in the Accuracy, which increased from 0.907 to 0.910, and the F1-Score, which increased from 0.893 to 0.897. The most prominent indicator of model quality improvement is reflected in the increase in the AUC value from 0.949 to 0.957, which indicates that the optimized model has superior generalization capabilities and inter-class separability in distinguishing review sentiments compared to the baseline model. Confusion matrix analysis showed that the most misclassifications occurred in neutral reviews, which are often categorized as positive. This occurs because many neutral reviews contain positive words but in ambiguous contexts, such as "the ride was fine, nothing special." In this case, the model tends to focus on the positive word "fine" without understanding the neutral nuance of the sentence. This condition strengthens the argument that TF-IDF is unable to capture semantic relationships between words and could be replaced by contextual embeddings in the future. From a business perspective, this finding is significant. Uber can implement CatBoost as a component of its real-time review monitoring system to improve

.

customer responsiveness. Negative reviews can be immediately prioritized for follow-up, for example, regarding driver delays, vehicle issues, or app errors. Conversely, positive reviews can be leveraged to identify service strengths that need to be maintained and even promoted in branding strategies. With >90% accuracy, the CatBoost model has proven reliable enough to support data-driven decision-making. While CatBoost's Native Text Support provides a streamlined workflow, the TF-IDF + GridSearchCV approach yielded a higher AUC of 0.957, confirming that manual weighting of sentiment-bearing terms provides more granular sensitivity for Uber's specific customer.

These results are consistent with previous research suggesting that boosting models, particularly CatBoost, outperform traditional methods such as Logistic Regression or Naive Bayes in large-scale text analysis. The experimental results confirm that CatBoost outperforms XGBoost and LightGBM in terms of stability and generalization, particularly when the dataset presents an imbalanced class distribution. By utilizing Ordered Boosting, the model achieved an AUC of 0.957, indicating a superior ability to distinguish between sentiment classes compared to default gradient boosting configurations, especially when the dataset has an imbalanced class distribution. A limitation of this research lies in the use of TF-IDF as a feature representation. Although effective, TF-IDF is only frequency-based and unable to capture semantic meaning. Therefore, further research can be directed at integrating deep contextual embeddings such as BERT or IndoBERT to improve the model's semantic understanding. Furthermore, the use of data balancing techniques such as SMOTE or ADASYN can also be explored to reduce bias due to class imbalance. Overall, this research confirms that the success of sentiment analysis is determined not only by the algorithm, but also by data quality, preprocessing techniques, feature representation, and hyperparameter optimization strategies. In this context, CatBoost with GridSearchCV proved to be the most effective combination for sentiment analysis on the Uber Customer Reviews Dataset 2024, as it produced the best balance between accuracy, precision, recall, and generalization ability.

4. CONCLUSION

This research demonstrates that the CatBoost algorithm, through its Ordered Boosting mechanism, provides a robust solution for sentiment analysis that minimizes overfitting better than rival algorithms like XGBoost and LightGBM, while the TF-IDF vectorization transforms text into numeric form, CatBoost's ability to handle heterogeneous data—including review length and thumbs-up counts—resulted in a significant AUC improvement to 0.957. The experimental results confirm that tuning the model using GridSearchCV led to consistent performance enhancements across all metrics, with Accuracy rising from 0.907 to 0.910 and the F1-Score increasing from 0.893 to 0.897. Most notably, the AUC score improved significantly from 0.949 to 0.957, indicating a superior ability to distinguish between sentiment classes compared to the default model. Despite these strong global metrics, a granular analysis via the confusion matrix exposes a distinct bias; while the model exhibits exceptional precision in identifying Positive and Negative sentiments, it failed to correctly classify Neutral instances, misinterpreting them as polar sentiments due to the likely ambiguity of the text and class imbalance. To overcome the limitations identified in this study and further enhance predictive capabilities, several avenues for future research are recommended. First, addressing the issue of class imbalance is paramount; future works should incorporate synthetic data augmentation techniques, such as SMOTE or ADASYN, to improve the model's learning of minority classes, like the Neutral sentiment. Second, to better capture the nuance and context of ambiguous reviews that caused the misclassification in this study, shifting from traditional feature extraction to advanced deep learning architectures, such as BERT or RoBERTa, is highly advised. Finally, to provide more actionable business insights for application developers, the scope of analysis could be expanded to Aspect-Based Sentiment Analysis (ABSA), allowing for the specific identification of service components, such as driver behavior or app functionality, that drive user satisfaction or dissatisfaction.

ACKNOWLEDGEMENT

We would like to express our gratitude to all parties who have contributed and fully supported in completing the research and writing of this paper completely. We would also like to express our gratitude to FMIPA UNNES which through the Lecturer Research program with Agreement Number: 45.21.4/UN37/PPK.04/2025 has provided very meaningful financial support in the implementation of this research so that our research can run smoothly and successfully.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

REFERENCES

[1] X. Xu, Y. Wang, Q. Zhu, and Y. Zhuang, "Time matters: Investigating the asymmetric reflection of online reviews on customer satisfaction and recommendation across temporal lenses," *Int J Inf Manage*, vol. 75, p. 102733, 2024, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102733.

[2] A. Boukis, L. Harris, and C. D. Koritos, "'Give me an upgrade or I will give you a bad review!' Investigating customer threats in the hospitality industry," *Tour Manag*, vol. 104, p. 104927, 2024, doi: https://doi.org/10.1016/j.tourman.2024.104927.

[3] M. Nilashi *et al.*, "The nexus between quality of customer relationship management systems and customers' satisfaction: Evidence from online customers' reviews," *Heliyon*, vol. 9, no. 11, p. e21828, 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e21828.

[4] T. Bruno *et al.*, "A blockchain-based platform for incentivizing customer reviews in the grocery industry," *Blockchain: Research and Applications*, vol. 5, no. 4, p. 100226, 2024, doi: https://doi.org/10.1016/j.bcra.2024.100226.

[5] A. H. Tahir, M. Adnan, and Z. Saeed, "The impact of brand image on customer satisfaction and brand loyalty: A systematic literature review," *Heliyon*, vol. 10, no. 16, p. e36254, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36254.

[6] F. Carichon, C. Ngouma, B. Liu, and G. Caporossi, "Objective and neutral summarization of customer reviews," *Expert Syst Appl*, vol. 255, p. 124449, 2024, doi: https://doi.org/10.1016/j.eswa.2024.124449.

[7] M. Cai and C. Yang, "Customer preference analysis integrating online reviews: An evidence theory-based method considering criteria interaction," *Eng Appl Artif Intell*, vol. 133, p. 108092, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108092.

[8] N. Wang, T. S. H. Teo, S. Liu, and V. K. G. Lim, "Hotel reviews during the pandemic: Encouraging repeat customers to 'speak up' through management response," *Int J Hosp Manag*, vol. 120, p. 103765, 2024, doi: https://doi.org/10.1016/j.ijhm.2024.103765.

[9] A. Amato, J. R. Osterrieder, and M. R. Machado, "How can artificial intelligence help customer intelligence for credit portfolio management? A systematic literature review," *International Journal of Information Management Data Insights*, vol. 4, no. 2, p. 100234, 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100234.

[10] S. Chen, Z. Xu, D. Xu, and X. Gou, "Customer purchase prediction in B2C e-business: A systematic review and future research agenda," *Expert Syst Appl*, vol. 252, p. 124261, 2024, doi: https://doi.org/10.1016/j.eswa.2024.124261.

[11] M. Zhai, X. Wang, and X. Zhao, "The importance of online customer reviews characteristics on remanufactured product sales: Evidence from the mobile phone market on Amazon.com," *Journal of Retailing and Consumer Services*, vol. 77, p. 103677, 2024, doi: https://doi.org/10.1016/j.jretconser.2023.103677.

[12] S. Soklaridis, A. M. Geske, and S. Kummer, "Key characteristics of perceived customer centricity in the passenger airline industry: A systematic literature review," *Journal of the Air Transport Research Society*, vol. 3, p. 100031, 2024, doi: https://doi.org/10.1016/j.jatrs.2024.100031.

[13] J. Langevin *et al.*, "Customer enrollment and participation in building demand management programs: A review of key factors," *Energy Build*, vol. 320, p. 114618, 2024, doi: https://doi.org/10.1016/j.enbuild.2024.114618.

[14] B. Burhanudin, "Managing social commerce: does customer review quality matter?," *Procedia Comput Sci*, vol. 234, pp. 1459–1466, 2024, doi: https://doi.org/10.1016/j.procs.2024.03.146.

[15] Y. Liu, T.-H. You, J. Zou, and B.-B. Cao, "Modelling customer requirement for mobile games based on online reviews using BW-CNN and S-Kano models," *Expert Syst Appl*, vol. 258, p. 125142, 2024, doi: https://doi.org/10.1016/j.eswa.2024.125142.

[16] D. Leocádio, L. Guedes, J. Oliveira, J. Reis, and N. Melão, "Customer Service with AI-Powered Human-Robot Collaboration (HRC): A Literature Review," *Procedia Comput Sci*, vol. 232, pp. 1222–1232, 2024, doi: https://doi.org/10.1016/j.procs.2024.01.120.

[17] H. Li, H. Liu, H. Hailey Shin, and H. Ji, "Impacts of user-generated images in online reviews on customer engagement: A panel data analysis," *Tour Manag*, vol. 101, p. 104855, 2024, doi: https://doi.org/10.1016/j.tourman.2023.104855.

[18] Y. A. Laghbi and M. Al Dhoayan, "Examining how customers perceive community pharmacies based on Google maps reviews: Multivariable and sentiment analysis," *Exploratory Research in Clinical and Social Pharmacy*, vol. 15, p. 100498, 2024, doi: https://doi.org/10.1016/j.rcsop.2024.100498.

[19] M. Zaman, C. C. Tan, M. S. Islam, and K. M. Selem, "Hospitality customer intentions to write fake online reviews: A cross-cultural approach," *Int J Hosp Manag*, vol. 120, p. 103775, 2024, doi: https://doi.org/10.1016/j.ijhm.2024.103775.

[20] L. Kim, T. Jindabot, and S. F. Yeo, "Understanding customer loyalty in banking industry: A systematic review and meta analysis," *Heliyon*, vol. 10, no. 17, p. e36619, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36619.

[21] S. Bellary, P. Kumar Bala, and S. Chakraborty, "Utilizing online reviews for analyzing digital healthcare consultation services: Examining perspectives of both healthcare customers and healthcare professionals," *Int J Med Inform*, vol. 191, p. 105587, 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105587.

[22] E. B. Firmansyah, M. R. Machado, and J. L. R. Moreira, "How can Artificial Intelligence (AI) be used to manage Customer Lifetime Value (CLV)—A systematic literature review," *International Journal of Information Management Data Insights*, vol. 4, no. 2, p. 100279, 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100279.

[23] M. T. H. Le, "Fostering product quality and Brand Trust by QR code traceability and customer reviews: The moderating role of brand reputation in Blockchain," *The Journal of High Technology Management Research*, vol. 35, no. 1, p. 100492, 2024, doi: https://doi.org/10.1016/j.hitech.2024.100492.

[24] T. Rahman, M. L. Othman, S. B. Mohd Noor, W. F. Binti Wan Ahmad, and M. F. Sulaima, "Methods and attributes for customer-centric dynamic electricity tariff design: A review," *Renewable and Sustainable Energy Reviews*, vol. 192, p. 114228, 2024, doi: https://doi.org/10.1016/j.rser.2023.114228.

[25] L. Zhang, Y. Xuan, Z. Li, P. Gao, and Y. Zheng, "How to obtain customer requirements for each stage of the product life cycle from online reviews: Using mobile phones as an example," *Journal of Retailing and Consumer Services*, vol. 80, p. 103928, 2024, doi: https://doi.org/10.1016/j.jretconser.2024.103928.

[26] T. Nguyen-Sy, "Optimized hybrid XGBoost-CatBoost model for enhanced prediction of concrete strength and reliability analysis using Monte Carlo simulations," *Appl Soft Comput*, vol. 167, p. 112490, 2024, doi: https://doi.org/10.1016/j.asoc.2024.112490.

[27] X. Huang, W. Liu, Q. Guo, and J. Tan, "Prediction method for the dynamic response of expressway lateritic soil subgrades on the basis of Bayesian optimization CatBoost," *Soil Dynamics and Earthquake Engineering*, vol. 186, p. 108943, 2024, doi: https://doi.org/10.1016/j.soildyn.2024.108943.

[28] X. Ren, H. Yu, X. Chen, Y. Tang, G. Wang, and X. Du, "Application of the CatBoost Model for Stirred Reactor State Monitoring Based on Vibration Signals," *CMES - Computer Modeling in Engineering and Sciences*, vol. 140, no. 1, pp. 647–663, 2024, doi: https://doi.org/10.32604/cmes.2024.048782.

.

[29] H. Qiu, Y. Xia, C. Xiang, F. Xu, L. Sun, and Y. Zou, "Prediction of hydrogen storage in metal-organic frameworks using CatBoost-based approach," *Int J Hydrogen Energy*, vol. 79, pp. 952–961, 2024, doi: https://doi.org/10.1016/j.ijhydene.2024.07.078.

[30] M. Zahid *et al.*, "Factors affecting injury severity in motorcycle crashes: Different age groups analysis using Catboost and SHAP techniques," *Traffic Inj Prev*, vol. 25, no. 3, pp. 472–481, 2024, doi: https://doi.org/10.1080/15389588.2023.2297168.

[31] E. Ghorbani and S. Yagiz, "Estimating the penetration rate of tunnel boring machines via gradient boosting algorithms," *Eng Appl Artif Intell*, vol. 136, p. 108985, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108985.

[32] Y. Li, Y. Duan, Y. Zhou, J. Yang, F. Li, and A. Yang, "Research on prediction model of iron ore powder sintering foundation characteristics based on FOA-Catboost algorithm," *Alexandria Engineering Journal*, vol. 86, pp. 603–615, 2024, doi: https://doi.org/10.1016/j.aej.2023.12.015.

[33] D. K. Singh and S. Kumar, "Techno-economics of high ash coal gasification: A machine learning approach using CatBoost model," *J Clean Prod*, vol. 481, p. 144160, 2024, doi: https://doi.org/10.1016/j.jclepro.2024.144160.

[34] M. Karbasi, M. Jamei, M. Ali, A. Malik, and Z. M. Yaseen, "Forecasting weekly reference evapotranspiration using Auto Encoder Decoder Bidirectional LSTM model hybridized with a Boruta-CatBoost input optimizer," *Comput Electron Agric*, vol. 198, p. 107121, 2022, doi: https://doi.org/10.1016/j.compag.2022.107121.

[35] Z. Ge *et al.*, "Quantifying and comparing the effects of key chemical descriptors on metal–organic frameworks water stability with CatBoost and SHAP," *Microchemical Journal*, vol. 196, p. 109625, 2024, doi: https://doi.org/10.1016/j.microc.2023.109625.

[36] J. Bian, J. Wang, and Q. Yece, "A novel study on power consumption of an HVAC system using CatBoost and AdaBoost algorithms combined with the metaheuristic algorithms," *Energy*, vol. 302, p. 131841, 2024, doi: https://doi.org/10.1016/j.energy.2024.131841.

[37] J. Yu, W. Zheng, L. Xu, F. Meng, J. Li, and L. Zhangzhong, "TPE-CatBoost: An adaptive model for soil moisture spatial estimation in the main maize-producing areas of China with multiple environment covariates," *J Hydrol (Amst)*, vol. 613, p. 128465, 2022, doi: https://doi.org/10.1016/j.jhydrol.2022.128465.

BIOGRAPHIES OF AUTHORS

Yahya Nur Ifriza is an Assistant Professor Information System Study Program at Universitas Negeri Semarang, Indonesia. He holds study in Informatics and Computer Engineering from Universitas Negeri Semarang Indonesia in 2014. He received a master's degree in information systems from Diponegoro University Indonesia in 2017 with the Thesis "Expert system irrigation management of agricultural reservoir system using analytical hierarchy process (AHP) and forward chaining method". In addition, she is serving as Head of Public Relations and Reputation Group (2022–present). His research interests include CRM, Social Marketing, Customer Behaviour, Machine Learning, Natural Language Processing and Recommendation System. He can be contacted at email: yahyanurifriza@mail.unnes.ac.id

Ratna Nur Mustika Sanusi is a lecturer in the Mathematics Study Program at Universitas Negeri Semarang, Indonesia. She completed her bachelor's degree in mathematics from Universitas Sebelas Maret in 2020, and earned her master's degree in Statistics and Data Science from IPB University in 2022 with a thesis titled "Simulation of the SARIMA Model with Three-Way ANOVA and Its Application in Forecasting Large Chilli Prices in Five Provinces on Java Island." Her research interests include simulation studies, forecasting, clustering, classification, sentiment analysis, and data-driven business applications. She can be contacted at rnmustika@mail.unnes.ac.id

Hendra Febriyanto is an Assistant Lecturer at Universitas Negeri Semarang, Indonesia. He holds a master's degree in biology education, with research interests in educational communication, ethnoscience, and biology learning. In addition to his academic work, he is actively involved in community engagement programs focused on enhancing communication skills among teachers and students. He is also the founder of firstwalk.id, a learning platform dedicated to communication and public speaking. Several of his works have been granted intellectual property rights (HaKI). Contact hendrafebri@mail.unnes.ac.id

Azlina Kamarudin is a Lecturer in the Department of Computer Science at Universiti Teknologi PETRONAS (UTP), which she joined in 2024. She earned her Ph.D. in Computer Science, specializing in Computer Networks, from Universiti Teknologi Malaysia (UTM) in 2021, following an M.Sc. in Electronic Systems Engineering from the University of Essex, United Kingdom. Before joining UTP, she served as a Lecturer at UTM from 2005 and gained professional experience as a freelance programmer. Her research interests include data communications, Internet of Things (IoT), and machine learning, with a focus on developing intelligent, connected systems and advancing innovations in networked technologies. She has contributed to multiple research projects, secured research grants, published in journals and conferences, delivered guest lectures at universities, and mentored students at both undergraduate and postgraduate levels.

.