ISSN 2723-6471

Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques

Pandu Pratama Putra[1,*], M. Khairul Anam[2], Andi Supriadi Chan[3], Abrar Hadi[4], Nofri Hendri[5], Alkadri masnur[6]

[1]*Department of Informatics Engineering, Universitas Lancang Kuning, Jl. Yos Sudarso, Pekanbaru and 28266, Indonesia*

[2]*Department of Informatics, Universitas Samudra, Jl, Prof. Dr. Syarief Thayeb, Langsa and 24416, Indonesia*

[3]*Department of Graphic Multimedia Engineering Technology, Politeknik Negeri Medan, Jl. Almamater, Medan and 20155, Indonesia*

[4]*Department of Informatics Management, Politeknik LP3I Kampus Padang, Jl. By Pass km.7, Padang and 25147, Indonesia*

[5,6]*Department of Educational Technology, Universitas Negeri Padang, Jl. Prof. Dr. Hamka, Padang and 25011, Indonesia*

(Received: December 13, 2024; Revised: January 11, 2025; Accepted: February 1, 2025; Available online: March 3, 2025)

Abstract

This research addresses the challenge of imbalanced sentiment classes in hotel review datasets obtained from Traveloka by integrating SMOTE (Synthetic Minority Oversampling Technique) with ensemble machine learning methods. The study aimed to enhance the classification of Positive, Negative, and Neutral sentiments in customer reviews. Data preprocessing techniques, including tokenization, stemming, and stopword removal, prepared the textual data for analysis. Various machine learning models—CART, KNN, Naive Bayes, and Random Forest—were evaluated individually and in ensemble configurations such as Bagging, Stacking, Soft Voting, and Hard Voting. The Stacking ensemble approach, utilizing Logistic Regression as a meta-classifier, demonstrated superior performance with an accuracy, precision, recall, and F1-score of 88%, outperforming Bagging (86%), Hard Voting (84%), and Soft Voting (81%). The findings highlight the effectiveness of SMOTE in balancing sentiment classes, particularly improving the classification of underrepresented Neutral and Negative categories. The novelty of this study lies in the comprehensive use of ensemble techniques combined with SMOTE, which significantly enhanced prediction stability and accuracy compared to previous approaches. These results provide valuable insights into leveraging advanced machine learning techniques for sentiment analysis, offering practical implications for improving customer experience and service quality in the hospitality industry.

Keywords: Sentiment Analysis, SMOTE, Ensemble Learning, Hotel Reviews

1. Introduction

Indonesia is one of the countries with significant tourism potential [1]. Its natural beauty, cultural diversity, and historical richness make the tourism sector one of the primary sources of national income [2]. Indonesian tourism has developed rapidly, with millions of local and international tourists visiting various destinations each year [3]. One crucial aspect contributing to the success of the tourism sector is the quality of hotel services. Hotels play an integral role in the tourist experience, where comfort, hospitality, and satisfying services are key factors that determine tourists' perceptions and satisfaction levels [4].

This study analyzes reviews from the Traveloka application, which is widely used by people to book hotels. Traveloka was chosen as the dataset source due to its extensive user base in Southeast Asia, particularly Indonesia, where it dominates the online travel market. This makes the dataset highly representative of regional customer behavior and preferences. Additionally, Traveloka provides a rich variety of reviews that include different sentiments, offering a balanced perspective on customer satisfaction and dissatisfaction. These reviews, typically submitted by tourists through online travel platforms like Traveloka, provide clear insights into real customer experiences. The reviews can

***Corresponding author: Pandu Pratama Putra (pandupratamaputra@unilak.ac.id)**
DOI: https://doi.org/10.47738/jads.v6i2.618

consist of positive comments reflecting satisfaction or criticisms indicating issues with the services provided [5]. With the advancement of technology, sentiment analysis of these reviews has become a valuable tool to help Traveloka's management understand customer perceptions more deeply [6]. Sentiment analysis enables Traveloka to identify trends in the reviews, whether they relate to services, facilities, or other factors connected to the guest experience or application use [7]. Consequently, hotels and Traveloka can take necessary actions to improve the quality of their services based on the extracted data.

Previous research on sentiment analysis using machine learning has been widely conducted. For instance, a study by [8] compared three algorithms for sentiment analysis: Support Vector Machine (SVM) achieved 68% accuracy, Naïve Bayes (NB) achieved 69%, and Random Forest achieved 66%. Another study analyzed hotel reviews in Indonesia and achieved relatively low accuracy using various machine learning algorithms: LR 54%, NB 53%, and SVM 54% [9]. This further highlights the challenge of class imbalance, as these models struggle to learn meaningful patterns from the underrepresented classes. Other research efforts have focused on improving machine learning performance using different methods. For example, a study by [10] enhanced the Light Gradient Boosting Machine (LGBM) algorithm by adding Bigram LDA Filter and Word2Vec, achieving an accuracy of 86.6%. Another study improved machine learning performance using ensemble methods, specifically stacking and voting, with the latter achieving an accuracy of 85% and the former 81%, by combining several base algorithms, namely Random Forest (RF), SVM, and XGBoost [11]. These findings underscore the persistent limitations of prior studies in addressing class imbalance, as well as the need for more comprehensive approaches that combine techniques like SMOTE and ensemble methods to enhance performance and ensure fair representation across all sentiment classes.

In sentiment analysis of customer reviews, especially when handling imbalanced data, the combination of Machine Learning (ML) and Natural Language Processing (NLP) becomes essential. NLP processes and understands customer review text through techniques such as tokenization, stemming, and lemmatization, making the review data usable for machine learning models [12]. After the NLP process, ML is applied to build a classification model capable of predicting the sentiment of the reviews—whether positive, negative, or neutral. This combination enables sentiment analysis to be conducted automatically on a large scale, which is useful for understanding customer perceptions without manual intervention [13]. One significant challenge in sentiment analysis is imbalanced data, where positive reviews often far outnumber negative reviews [14]. To address this, SMOTE (Synthetic Minority Over-sampling Technique) is used. SMOTE helps balance the dataset by synthesizing new samples for minority classes (e.g., negative reviews), allowing machine learning models to learn more effectively from both classes [15]. With SMOTE, models become more sensitive to underrepresented classes, improving overall performance, especially in detecting negative reviews crucial for service improvement.

Commonly used machine learning algorithms for sentiment analysis include Naïve Bayes, SVM, Random Forest, and Logistic Regression. Naïve Bayes is chosen for its efficiency in handling text [16], while SVM excels with high-dimensional data and produces robust classifications [17]. Random Forest is effective in reducing overfitting and delivering stable results [18], and Logistic Regression provides predictions that are easy to interpret and understand [19]. To further improve model performance, ensemble machine learning techniques such as bagging, voting, and stacking are used. Bagging reduces variance by training multiple models on different subsets of the data [20]. Stacking combines multiple models using a meta-model to produce final predictions with higher accuracy [21]. Voting, on the other hand, combines the predictions of several base models to generate an optimal final solution [22]. By incorporating SMOTE into this process, the resulting models become not only more accurate but also more responsive to imbalanced data, ensuring that rare negative reviews are still detected effectively. The combination of SMOTE and ensemble machine learning produces stronger and more effective sentiment analysis models, improving sentiment prediction quality and providing valuable insights for companies to enhance their services.

2. Related Studies

The research conducted by [23] discusses the use of the Naive Bayes algorithm combined with SMOTE to address data imbalance in customer sentiment analysis from Twitter reviews. The goal is to automate customer sentiment classification, allowing TokopediaCare to respond more quickly. The results show that applying SMOTE significantly

improves the model's accuracy compared to using Naive Bayes alone. This approach accelerates sentiment analysis while ensuring more accurate results in handling various types of customer sentiments.

Another study by [24] focuses on implementing ensemble methods to analyze sentiment in Google Play Store reviews, specifically for applications such as Zoom and Shopee. The study aims to compare the effectiveness of ensemble models like Random Forest and Boosting with individual algorithms such as Naive Bayes and SVM. The research begins with a preprocessing step, including data cleaning, tokenization, and stopword removal to prepare the text reviews before classification. The results show that ensemble models, particularly Random Forest, deliver superior performance in sentiment classification, achieving 94.15% accuracy for Zoom reviews and 80.69% for Shopee reviews. These findings indicate that ensemble approaches are more effective in handling the complexity and variability of review data. The authors also note that while ensemble models improve performance, further development is needed, especially in addressing data imbalance and capturing more complex language nuances. Recommendations for future research include the use of Deep Learning techniques and testing models in other domains.

Furthermore, [25] analyze sentiment in user reviews of Shopee in Indonesia, focusing on public sentiment during the 11.11 Flash Sale event. The study uses TF-IDF (Term Frequency-Inverse Document Frequency) to calculate word weights in the reviews, and the results show a model accuracy of 90.76% under a 60:40 data split scenario. The study emphasizes the importance of understanding consumer sentiment to enhance strategies for events such as Flash Sales and improve customer service. Additionally, the findings show that the model effectively identifies negative sentiment during significant events, offering recommendations for Shopee to enhance marketing strategies and responsiveness to technical issues on the application.

The research by [26] focuses on sentiment analysis of customer reviews for the Ralali.com application on Google Play Store. In this study, three classification algorithms NB, SVM, and k-NN are compared to determine their effectiveness in classifying sentiment reviews. To address class imbalance in the dataset, the SMOTE method is applied. The results demonstrate that the Naive Bayes algorithm with SMOTE outperforms models without SMOTE in terms of recall and precision. However, some challenges remain in maintaining accuracy because SMOTE slightly reduces performance for the majority class. The final results indicate that combining SMOTE with Naive Bayes is the most effective approach for handling imbalanced data, particularly for minority class classification. The study also highlights the importance of two-step data preprocessing, including data cleaning and stemming, which significantly improves the quality of the models used. This research provides recommendations for Ralali.com developers to improve systems based on user feedback identified through sentiment analysis.

Other Research [27] focuses on using ensemble methods that combine Naive Bayes, Decision Trees, Multilayer Perceptron, and Logistic Regression to improve sentiment analysis accuracy on Twitter. This study highlights the challenges of analyzing social media data, such as informal language and ambiguous expressions. The results show that the proposed ensemble classifier outperforms individual models in classifying positive and negative sentiments, with Naive Bayes being the best-performing individual model.

Subsequent research by [28] This research focuses on the use of ensemble methods combining Naive Bayes, Decision Trees, Multilayer Perceptron, and Logistic Regression to improve sentiment analysis accuracy on Twitter. The study highlights the challenges of analyzing social media data, such as informal language and ambiguous expressions. The results indicate that the proposed ensemble classifier outperforms individual models in classifying positive and negative sentiments, with Naive Bayes being the best-performing model among the individual models.

Previous research predominantly employed classical machine learning methods such as Naive Bayes, Random Forest, and Support Vector Machine for sentiment analysis. Several studies combined the SMOTE method to handle class imbalance in sentiment data, which led to improved classification performance. In addition, some studies utilized ensemble techniques such as boosting or majority voting to enhance sentiment prediction accuracy.

Compared to previous studies, this research adopts a more structured approach by involving various models such as Naive Bayes, Random Forest, KNN, and CART and preprocessing techniques such as TF-IDF and SMOTE to address data imbalance. Voting and stacking with Logistic Regression are used to improve predictive performance, and the final results are thoroughly evaluated using relevant metrics. Unlike previous studies that tended to rely on a single

model or optimization technique, this study is more comprehensive, implementing a combination of models and more diverse ensemble techniques. These include Bagging with Random Forest, stacking with base algorithms such as NB, RF, KNN, and CART with Logistic Regression as the meta-learner, as well as Voting using both soft and hard voting approaches with the same base algorithms. Additionally, this study also evaluates baseline algorithms, ensuring a more holistic approach to sentiment analysis.

3. Methodology

The methodology applied in this research is illustrated in figure 1, which provides an overview of the research workflow. This diagram outlines the sequential steps involved in the process, starting from data collection and preprocessing, followed by the application of SMOTE to address class imbalance, machine learning model training, and performance evaluation. Each step is designed to ensure the effectiveness of the sentiment classification, particularly in handling imbalanced data. Figure 1 serves as a visual guide to help readers understand the interconnected stages of the methodology.

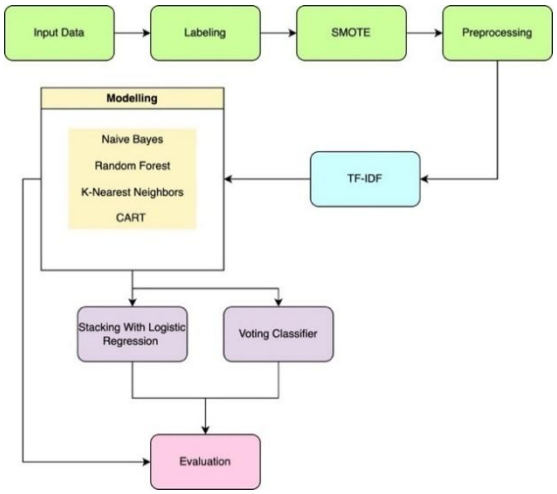


Figure 1. Overview of the Research flow

3.1. Input Data

The initial stage of this research involved collecting raw data to serve as the foundation for subsequent analysis. The dataset was sourced from Kaggle, a popular platform for sharing datasets, specifically from the following link: https://www.kaggle.com/datasets/mgustiansyah/traveloka-id-application-rating-and-review-dataset. This dataset contains a compilation of user reviews and ratings submitted by customers of the Traveloka application. These reviews are valuable as they provide insights into customer satisfaction, preferences, and potential issues faced by users of the platform. The dataset is rich in textual and numerical data, allowing for comprehensive analysis, including sentiment analysis and NLP tasks. By utilizing this dataset, the study aims to explore patterns and trends that can contribute to improving the Traveloka application's user experience and overall functionality. The inclusion of real-world data ensures that the findings are applicable and relevant to practical scenarios.

3.2. Labelling

Traveloka application was directly labeled based on sentiment indicators embedded within the data, such as star ratings or explicit positive and negative expressions in the reviews. For instance, reviews with higher ratings and positive language were labeled as “positive”, while those with lower ratings and negative language were labeled as “negative”. Reviews that did not convey a strong sentiment or had neutral ratings were labeled as “neutral”.

The second stage of labeling involved input and validation from the General Manager (GM) of a hotel located in Pekanbaru. As an expert in hospitality and customer feedback analysis, the GM provided valuable insights to refine the labels, ensuring the sentiment categorization aligned with the real-world understanding of customer satisfaction and dissatisfaction. This manual validation step ensured that the labeling process captured nuanced expressions of sentiment that automated methods might overlook, enhancing the overall quality and reliability of the labeled dataset.

3.3. SMOTE

SMOTE is a widely used technique to address imbalanced datasets by generating synthetic samples for underrepresented classes. This approach involves identifying k-nearest neighbors of data points in the minority class and creating new synthetic samples along the line segments between these neighbors. Unlike traditional oversampling methods that duplicate existing instances, SMOTE generates new, unique samples, which reduces the risk of overfitting [29].

In this study, SMOTE was employed to balance the sentiment classes Positive, Negative, and Neutral, within the dataset. The initial dataset showed significant class imbalances, where the Positive class was dominant, and the Neutral and Negative classes were underrepresented. By applying SMOTE, synthetic samples were generated for the Negative and Neutral classes to equalize their representation with the Positive class. This preprocessing step ensured that the machine learning models trained on the dataset could learn features from all sentiment classes effectively, leading to more robust and unbiased predictions.

The implementation of SMOTE significantly improved the model's ability to classify minority classes accurately, especially in scenarios where the raw dataset would have otherwise caused the model to bias its predictions toward the majority class. By balancing the dataset, the study ensured a fair evaluation of the model's performance across all sentiment categories, ultimately leading to a more reliable and inclusive sentiment classification system.

3.4. Preprocessing

In this research, the preprocessing step plays an important role in preparing text data for machine learning models. In figure 2 is the initial data taken from Kaggle.

	Nama	Bintang	Tanggal dan Waktu	Ulasan
0	Edy Suranto	5.0	11/17/2023 9:33	Aplikasi yang sangat bagus, banyak sekali prom...
1	p mustika	2.0	11/16/2023 21:27	Semua aman terkendali sampai akun paylater di ...
2	Akhmad Jailani	5.0	11/9/2023 5:14	Aplikasi yang sangat bagus, banyak promo yang ...
3	Laras Ayu	5.0	11/13/2023 9:07	Aplikasi yang sangat bagus, banyak promo yang ...
4	19 chanel	5.0	11/17/2023 5:29	Aplikasi ini sangat memudah urusana perjalanan...
...	...	...	...	...
63477	Angga	5.0	1/27/2021 12:46	bagus kok
63478	Haris Adiyatma	5.0	1/27/2021 5:53	Good Job
63479	fadli adonis	5.0	1/27/2021 18:17	sangat bagus
63480	ADY STAR Production	5.0	1/29/2021 13:19	good promo
63481	Roni Bakka lamu official	5.0	1/24/2021 8:38	mantappp traveloka

Figure 2. Initial Data

Figure 2 illustrates the raw dataset containing information about user names, star ratings, review timestamps, and review content related to the Traveloka application. While this dataset provides an initial overview of user feedback, preprocessing is essential to ensure the quality and relevance of the data for further analysis. Preprocessing is necessary because raw data often contains irrelevant or potentially disruptive elements, such as user names and timestamps, which may not directly contribute to sentiment analysis. Additionally, review content may include unnecessary characters or symbols that need to be cleaned to facilitate more accurate analysis. This process also reduces the risk of bias caused by duplicate entries or inconsistencies in data formatting. By cleaning and normalizing the data, preprocessing aims to produce a more structured and ready-to-use dataset, ensuring that machine learning models can optimally identify critical patterns. It also allows for managing more relevant information, enhancing the accuracy and efficiency of sentiment analysis on user reviews. The following is the preprocessing process in this study, The preprocessing steps included tokenization, stemming, stopword removal, lowercasing, and punctuation removal. During tokenization, the text was split into individual words using the NLTK tokenizer, which ensures accurate word segmentation, including handling special characters and punctuation [30]. Stemming was applied using the Porter Stemming Algorithm to reduce words to their root forms, allowing the model to treat word variations like "running" and "ran" as the same feature. This approach is computationally efficient and widely used for English text [31]. Stopword removal involved utilizing a predefined list of stopwords from the NLTK library to eliminate frequently occurring words such as "and," "the," and "is" that do not add significant meaning to sentiment classification. Additionally, custom stopwords relevant to the dataset, such as domain-specific terms, were reviewed and adjusted [32]. All text was converted to lowercase to standardize the data and prevent the model from treating words like "Hotel" and "hotel" as separate features [33].

Lastly, unnecessary symbols, including punctuation marks, were removed to ensure the text was clean and ready for analysis [34].

These preprocessing steps ensure that the text data is structured consistently and appropriately for machine learning models, thus improving their ability to learn meaningful patterns. By specifying the tools and methods used, the preprocessing procedure becomes more transparent and reproducible. Figure 3 is the result of the preprocessing performed.

	Ulasan	Label
0	aplikasi bagus promo ditawarkanbanyak pilih de...	Positif
1	aman kendali akun paylater beku aktif tdk tung...	Negatif
2	aplikasi bagus promo tawar pilih destinasi wis...	Positif
3	aplikasi bagus promo tawar pilih destinasi wis...	Positif
4	aplikasi mudah urusana jalan kualitas percaya ...	Positif
...	...	...
63477	bagus	Positif
63478	good job	Positif
63479	bagus	Positif
63480	good promo	Positif
63481	mantappp traveloka	Positif

Figure 3. Preprocessing Result

Figure 3 illustrates the results of the preprocessing stage, showcasing a dataset that has been refined and standardized to enhance its usability for further analysis. Irrelevant columns such as names and timestamps have been removed, while review content has been processed to eliminate unnecessary symbols, punctuation, and inconsistencies. Additionally, the text has been tokenized, stemmed, and lowercased to ensure uniformity, and stopwords have been removed to focus on meaningful words. This preprocessing ensures that the dataset is structured, noise-free, and ready for machine learning models to effectively extract patterns and insights, particularly for sentiment analysis.

3.5. TF-IDF

In this study, TF-IDF is utilized to transform textual reviews from the dataset into numerical representations that can be effectively processed by machine learning algorithms. The method involves calculating the Term Frequency (TF) for each word, which reflects how frequently a term appears in a specific review relative to the total number of terms in that review. Simultaneously, the Inverse Document Frequency (IDF) assigns lower weights to terms that occur commonly across the dataset, ensuring that frequent but less informative words contribute minimally to the numerical representation. By combining TF and IDF, the resulting TF-IDF score emphasizes terms that are unique and significant to individual reviews, allowing the model to prioritize meaningful features over redundant or irrelevant ones. This preprocessing step plays a vital role in ensuring that the classification model focuses on critical aspects of the reviews, ultimately enhancing the accuracy of sentiment classification [35].

3.6. Modelling

The selection of machine learning models in this study was guided by their unique strengths and suitability for sentiment classification tasks. Each model offers specific advantages that address different aspects of the dataset and prediction objectives. The sentiment classification process utilized multiple algorithms, each with distinct strengths and characteristics. Naive Bayes was chosen for its simplicity and efficiency, particularly suitable for text analysis due to its probabilistic approach. This algorithm effectively classifies sentiment by leveraging the probabilities of word occurrences in positive, negative, and neutral classes, making it especially useful for large datasets where interpretability and speed are critical [16]. Random Forest was selected for its ability to address overfitting issues commonly associated with single decision trees. As an ensemble learning algorithm, Random Forest combines predictions from multiple decision trees using the bagging technique, ensuring robust and stable predictions even in datasets with complex patterns or noise [36].

K-Nearest Neighbors (KNN) provides a straightforward approach to classification by examining the majority label among the closest neighbors in the feature space. This algorithm excels in small to medium-sized datasets, where computational efficiency and local pattern recognition are essential [37]. Lastly, CART (Classification and Regression Trees) builds interpretable decision tree models capable of handling both classification and regression tasks. Its

flexibility in capturing non-linear relationships between features and outcomes makes it a valuable tool for sentiment classification, especially when working with datasets that exhibit varied feature distributions [38].

Each model was implemented to process the preprocessed sentiment data, generating predictions based on the labeled reviews. The diverse selection of algorithms ensures that the strengths of each approach are utilized, contributing to comprehensive and reliable sentiment analysis outcomes.

3.7. Stacking With Logistic Regression

Stacking is an ensemble technique that combines multiple models (Naive Bayes, Random Forest, KNN, CART) using Logistic Regression as the meta-model. The meta-model generates the final prediction based on the outputs of the base models. This technique aims to combine the strengths of multiple models to produce better and more accurate predictions than a single model [39].

3.8. Voting

In a Voting Classifier, predictions from multiple models are combined using either Hard Voting or Soft Voting mechanisms. In this approach, algorithms such as NB, RF, KNN, and Classification and Regression Tree (CART) are often used as base models due to their ability to handle diverse data characteristics. In Hard Voting, each model provides discrete class predictions, and the class with the most votes is chosen as the final result. This method is particularly effective in scenarios where individual models perform well and their predictions align frequently [40]. On the other hand, Soft Voting combines predictions by averaging the probabilities for each class from all models and selects the class with the highest average probability as the final prediction. While Soft Voting offers a probabilistic approach and can leverage the confidence scores of the models, it may perform worse than Hard Voting in certain cases. This is because Soft Voting's reliance on probability averaging can dilute the influence of strong models if other models provide less accurate or overly confident probability estimates [41]. Therefore, the choice between Hard Voting and Soft Voting should consider the characteristics, performance consistency, and contributions of base models such as NB, RF, KNN, and CART.

3.9. Evaluation

To evaluate the performance of the machine learning models, the study uses four key metrics: accuracy, precision, recall, and F1-score. Each metric provides distinct insights into the model's ability to classify data correctly, especially in scenarios where class imbalance may affect the results.

Accuracy: Accuracy is used to measure the proportion of total correct predictions out of all predictions made by the model. It provides an overall assessment of model performance [42]. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{True Positives (TP)+True Negative (TN)}}{\text{Total Number of Instances}} \quad (1)$$

This metric is particularly useful when the dataset is balanced.

Precision: Precision focuses on the model's ability to correctly predict positive cases and avoids labeling negative cases as positive. It is essential in situations where false positives are costly [43]. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positive (TP)+False Positives (FP)}} \quad (2)$$

High precision indicates that the model’s positive predictions are highly reliable.

Recall: Recall measures the ability of the model to identify all actual positive cases. This is crucial when false negatives carry a significant cost, such as in medical diagnoses [44]. The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}} \quad (3)$$

F1-score: The F1-score is the harmonic mean of precision and recall. It balances these two metrics to provide a comprehensive measure of model performance, especially in imbalanced datasets [45]. The formula for the F1-score is:

 ISSN 2723-6471

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (4)$$

A high F1-Score reflects a balance between precision and recall, indicating robust model performance.

These metrics are chosen to ensure a comprehensive evaluation of the models, addressing both their ability to make correct predictions overall (accuracy) and their sensitivity and specificity when dealing with imbalanced data (precision, recall, and F1-score). This holistic evaluation enables a more nuanced understanding of model behavior under varying conditions.

4. Results and Discussion

4.1. Result

The first step involves balancing the classes, as shown in figure 4, which illustrates the class distribution before balancing using SMOTE. The review data distribution appears imbalanced, with the "Positive" category containing the largest number of data points, followed by "Negative," while "Neutral" has the least. This imbalance can bias machine learning models, making it easier for the model to identify the "Positive" class while struggling to recognize the underrepresented "Neutral" class. Figure 5 shows the class distribution after balancing the data using SMOTE.

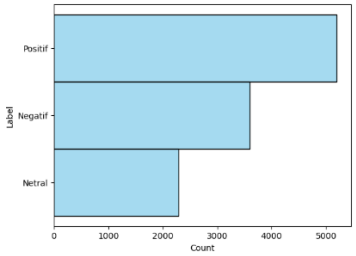


Figure 4. Data before balancing

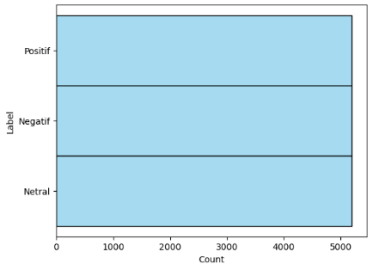
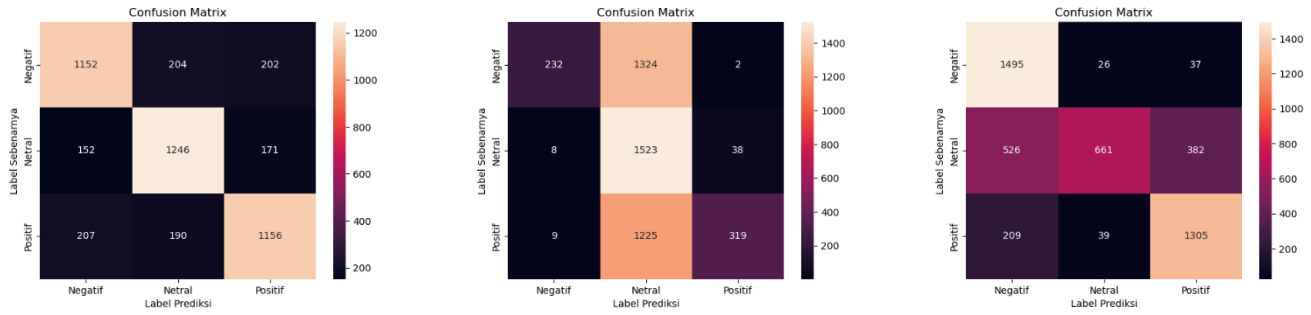


Figure 5. Data After balancing

Once SMOTE is applied, the data in each class ("Positive," "Negative," "Neutral") becomes balanced. SMOTE generates synthetic samples for the minority classes (Negative and Neutral), making the distribution equal to the "Positive" class. With balanced data, machine learning models can better identify patterns in all classes without bias caused by class imbalance. The next step involves testing the models using base algorithms, as shown in figure 6, which presents the confusion matrices for the following algorithms:



(a) CART (b) KNN (c) MNB

Figure 6. Confusion Matrices of Individual Algorithms (CART, KNN, and MNB)

For the CART model, the confusion matrix shows some classification errors across all classes. For instance, the "Negative" class is misclassified as "Neutral" 204 times and as "Positive" 202 times. Similarly, the "Neutral" class is often misclassified as "Positive," with 171 incorrect predictions. This indicates that CART struggles to accurately distinguish between the "Neutral" and "Positive" classes.

The KNN model shows improvements over CART, particularly in classifying the "Neutral" class. However, there are still errors, such as the "Positive" class being misclassified as "Neutral" 319 times. Despite this, KNN demonstrates better overall performance in identifying sentiment patterns compared to CART.

 ISSN 2723-6471

For the MNB model, the confusion matrix highlights strong performance in classifying the "Negative" and "Positive" classes with relatively fewer errors. However, the "Neutral" class remains challenging, as many predictions for this class are misclassified as "Negative" (661) and "Positive" (382). Overall, MNB performs well but requires further improvements to handle the "Neutral" class more effectively. The evaluation metrics for these models are summarized in table 1.

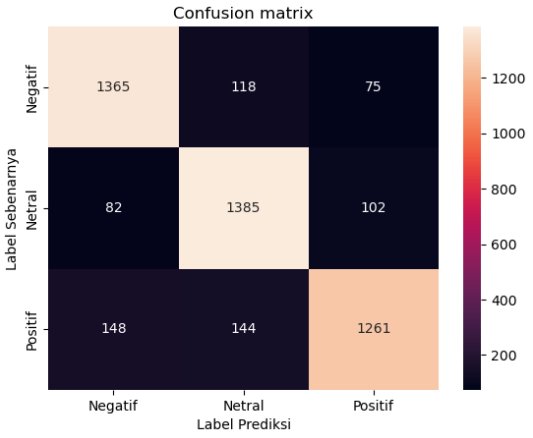
Table 1. Evaluation of Algorithm Based Metrics

Algorithm	Accuracy	Precision	Recall	F1-score
CART	76	76	76	76
KNN	44	73	44	38
MNB	74	78	74	72

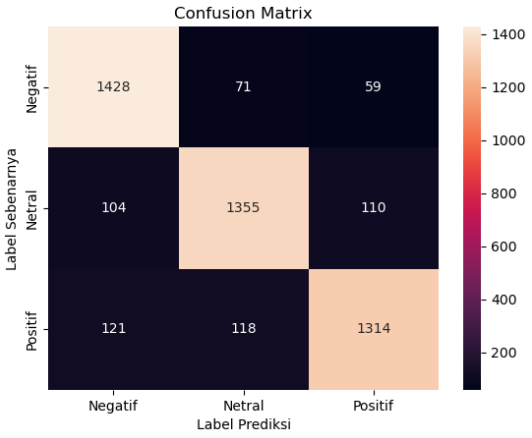
Table 1 summarizes the performance of three machine learning algorithms CART, KNN, and MNB (Multinomial Naive Bayes)—based on four evaluation metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. Precision focuses on the proportion of correctly predicted positive instances among all instances predicted as positive, while recall measures the proportion of actual positive instances correctly identified by the model. F1-score provides a balanced evaluation by combining precision and recall, especially useful when dealing with imbalanced datasets.

In the table, CART demonstrates consistent performance across all metrics, scoring 76%. This indicates that CART provides stable predictions but may struggle with nuanced differences in sentiment classes. KNN shows the lowest accuracy at 44%, indicating difficulty in classifying data accurately. Although its precision is relatively high at 73%, reflecting fewer false positives, its recall (44%) and F1-score (38%) are significantly lower, highlighting its inability to capture most positive cases. This performance is likely due to KNN's sensitivity to data distribution and noise.

MNB performs better than KNN, achieving an accuracy of 74% and the highest precision at 78%. Its recall (74%) and F1-score (72%) are also relatively strong, showing that MNB balances its predictions effectively. However, it still falls slightly short of CART in overall consistency. By combining these metrics, the table 1 shows that CART provides the most stable results, KNN struggles due to its sensitivity to data distribution, and MNB strikes a good balance between precision and recall, making it competitive for sentiment classification. The models were then evaluated using ensemble techniques, as shown in figure 7, which presents the confusion matrices for the following methods:

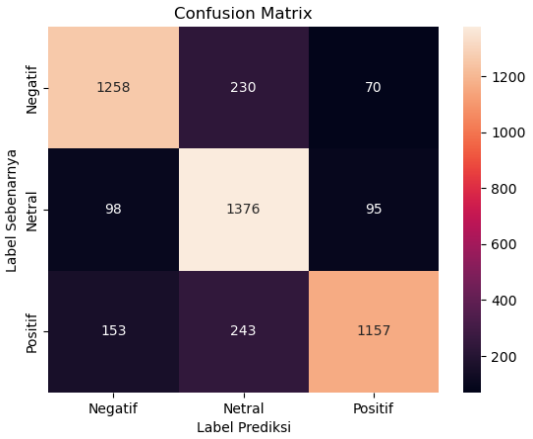


(a) Bagging: Random Forest (RF)

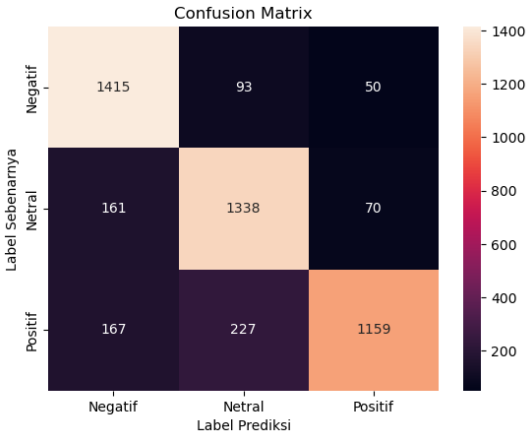


(b) Stacking: CART, KNN, SVM, RF: Logistic Regression

 ISSN 2723-6471



(c) Soft Voting: CART, KNN, SVM, RF



(d) Hard Voting: CART, KNN, SVM, RF

Figure 7. Confusion Matrices of Ensemble Techniques (Bagging, Stacking, Soft Voting, and Hard Voting)

The Bagging method with Random Forest produces stable performance, particularly in the "Negative" class, but still struggles to distinguish between the "Neutral" and "Positive" classes. This can be attributed to the overlap in class features, which affects the voting mechanism across multiple decision trees, even though Random Forest is well-suited for handling noisy data. Stacking, which combines predictions from base models (CART, KNN, SVM, RF) using Logistic Regression as a meta-classifier, outperforms the other methods. It handles class ambiguity more effectively due to the meta-classifier's ability to aggregate diverse prediction patterns, resulting in improved stability in the "Neutral" class despite minor errors between "Neutral" and "Positive" predictions.

Soft Voting, which averages prediction probabilities from the base models, performs evenly across all classes, with fewer errors in the "Negative" and "Positive" classes. However, the "Neutral" class shows higher misclassification rates, likely because averaging probabilities dilutes the influence of more accurate base models. Hard Voting, which uses majority voting from base models, performs well for the "Negative" class but exhibits significant errors in the "Neutral" class, where data is often misclassified as "Negative" or "Positive." This limitation arises because Hard Voting does not consider prediction probabilities, making it less flexible than Soft Voting. For models like KNN, lower recall and precision rates stem from its sensitivity to noisy data and reliance on neighbors, which can lead to errors when class boundaries are unclear or when classes are closely grouped. The evaluation metrics for ensemble methods are summarized in table 2.

Table 2. Evaluation of Ensemble Machine Learning

Algorithm	Accuracy	Precision	Recall	F1-score
Bagging: Random Forest (RF)	86	86	86	86
Stacking: CART, KNN, SVM, RF: Logistic Regression	88	88	88	88
Soft Voting: CART, KNN, SVM, RF	81	82	81	81
Hard Voting: CART, KNN, SVM, RF	84	84	84	84

Table 2 evaluates the performance of four ensemble machine learning methods—Bagging with RF, Stacking, Soft Voting, and Hard Voting—using four key metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions by dividing the number of correct predictions by the total instances. Precision focuses on the proportion of true positive predictions out of all predicted positives, which is important for minimizing false positives. Recall assesses the model's ability to identify actual positive instances, reflecting its sensitivity. F1-score, the harmonic mean of precision and recall, provides a balanced evaluation, particularly valuable when dealing with imbalanced datasets.

The results show that Stacking achieves the highest scores across all metrics (88%), demonstrating its ability to effectively combine predictions from multiple base models with Logistic Regression as the meta-classifier. Bagging

ISSN 2723-6471

with Random Forest follows closely, achieving consistent scores of 86% across all metrics, showcasing its robustness and stability. Hard Voting performs well with 84% across all metrics but lacks the flexibility of probability-based predictions. Soft Voting shows slightly lower performance (81% accuracy), attributed to its reliance on averaging probabilities, which may dilute the contributions of more accurate models. These metrics highlight the strengths and limitations of each ensemble method, emphasizing the superiority of Stacking in this analysis.

4.2. Discussion

The first step involves addressing data imbalance, as shown in Figure 4, which displays the class distribution before balancing using SMOTE. The data reveals that the "Positive" class has the highest count, followed by "Negative," while "Neutral" has the least. This imbalance can lead to bias in machine learning models, where they may struggle to identify minority classes like "Neutral." After applying SMOTE, as shown in Figure 5, the class distribution becomes balanced, ensuring that the model can learn patterns from all classes more effectively.

The next step is testing the models using individual algorithms, as shown in Figure 6. The three algorithms used are CART, KNN, and MNB. The confusion matrix for CART shows difficulties in distinguishing the "Neutral" and "Positive" classes, with some "Negative" data also misclassified. In contrast, KNN demonstrates improved classification of the "Neutral" class, but errors persist, particularly where "Positive" data is predicted as "Neutral." This poor performance in classifying the "Neutral" class can be attributed to the ambiguous nature of reviews in this category, which often contain mixed sentiments or lack clear indicators for classification. Additionally, models like KNN are sensitive to noisy or overlapping data, which exacerbates errors in distinguishing the "Neutral" class. MNB performs relatively well in classifying the "Negative" and "Positive" classes, but still struggles with the "Neutral" class, as evidenced by frequent misclassifications into other classes. This limitation may stem from the assumptions of the MNB algorithm, which treats features as independent, potentially oversimplifying complex textual patterns in the "Neutral" class. According to table 1, CART achieves a stable performance across accuracy, precision, recall, and F1-score, all at 76%, though it struggles to recognize minority classes. KNN has the lowest accuracy at 44%, with a higher precision of 73%, but its low recall indicates weaknesses in detecting all instances within a class. MNB performs better than KNN, achieving 74% accuracy and the highest precision at 78%, with balanced recall and F1-score values.

To further enhance classification performance, ensemble methods were applied, as illustrated in Figure 7, which displays the confusion matrices for four ensemble techniques: Bagging (Random Forest), Stacking, Soft Voting, and Hard Voting. Stacking, which combines base models (CART, KNN, SVM, RF) with Logistic Regression as a meta-classifier, delivers the best performance with high accuracy and stability in handling class ambiguity. Stacking outperforms other techniques because it leverages the strengths of diverse base models while mitigating their individual weaknesses through the meta-classifier. The Logistic Regression meta-classifier effectively identifies patterns in the combined outputs of base models, resulting in more accurate predictions. Bagging with Random Forest ranks second, achieving 86% accuracy, with stable performance but challenges in distinguishing the "Neutral" and "Positive" classes due to its reliance on multiple decision trees, which can dilute the model's ability to handle minority class boundaries. Soft Voting, which averages prediction probabilities from base models, provides balanced performance across all classes but has lower accuracy compared to Stacking and Bagging. The difficulty in classifying the "Neutral" class in Soft and Hard Voting is likely due to their reliance on aggregation techniques, which may overlook subtle differences in sentiment within ambiguous reviews. Hard Voting shows slightly better performance than Soft Voting but lacks flexibility as it relies only on majority voting, ignoring probability confidence.

According to table 2, Stacking achieves the best performance with an accuracy, precision, recall, and F1-score of 88%, followed by Bagging at 86%, Hard Voting at 84%, and Soft Voting at 81%. These results highlight that Stacking effectively combines the strengths of base models to produce accurate and stable predictions. Bagging with Random Forest also offers strong and reliable results, particularly in noisy data scenarios. Hard Voting and Soft Voting have limitations in handling complex data, especially for minority classes. The performance of these models is influenced by their respective hyperparameters. For instance, the number of neighbors in KNN, the tree depth in CART and Random Forest, and the regularization parameter in Logistic Regression were set using default values for initial testing. Although no advanced hyperparameter tuning was applied in this study, techniques such as Grid Search or Optuna could further optimize the models, potentially improving their performance and stability.

In conclusion, applying SMOTE successfully addressed the data imbalance issue, followed by testing individual and ensemble-based algorithms. Stacking emerged as the best method for sentiment classification due to its ability to combine base models effectively, thus improving overall performance. The challenges in classifying the "Neutral" class across all models indicate that the nature of "Neutral" reviews, often containing mixed or vague sentiments, poses significant challenges for machine learning models. Additionally, the models' underlying assumptions and aggregation mechanisms may not fully capture the subtleties of such ambiguous text data. The choice of ensemble technique depends on the specific needs and complexity of the data. A comparison with previous studies, as shown in table 3, demonstrates that this research outperforms earlier works using ensemble methods, achieving the highest accuracy of 88% with Stacking (CART, KNN, SVM, RF) and Logistic Regression as the meta-learner.

Table 3. Comparison with Previous Research

Researcher	Algorithm	Ensemble Technique	Accuracy
[22]	LR, KNN, SVM, DT, and LR	Soft Voting	78%
[46]	RF	Bagging	82%
[47]	MNB and SVM	Soft Voting	84%
[48]	LR, RF, and SVM with LR meta learner	Stacking	86%
[49]	XGBoost and RF with LSTM meta learner	Stacking	86%
This Research	CART, KNN, SVM, RF with LR meta learner	Stacking	88%

Table 3 compares several studies that implement machine learning algorithms and ensemble techniques for sentiment analysis, along with their achieved accuracy. Researchers [22] used a combination of Logistic Regression, KNN, SVM, Decision Tree, and Random Forest with the Soft Voting technique, resulting in an accuracy of 78%. Meanwhile, [46] applied Random Forest with the Bagging method, achieving a higher accuracy of 82%.

In contrast, [47] employed Soft Voting using Multinomial Naive Bayes and SVM, improving the accuracy to 84%. A more advanced ensemble technique, Stacking, was utilized by [48], combining Logistic Regression, Random Forest, and SVM, with Logistic Regression as the meta-learner, achieving an accuracy of 86%. Similarly, [49] used Stacking with XGBoost and RF with LSTM meta learner combined with LSTM as the meta-learner, resulting in a slightly improved accuracy of 86%.

Finally, this research outperformed all previous studies by achieving the highest accuracy of 88% using the Stacking method. This approach involved base models CART, KNN, SVM, and Random Forest, with Logistic Regression serving as the meta-learner. These results highlight the effectiveness of combining multiple models and advanced ensemble techniques to optimize performance in sentiment analysis.

While the results demonstrate that Stacking achieved the best accuracy among the ensemble methods, it is important to acknowledge the limitations encountered during its application. One specific challenge was the increased computational complexity and longer training times due to the involvement of multiple base models and a meta-classifier. Additionally, the performance of Stacking heavily relies on the selection and tuning of base models, as well as the meta-classifier, which might not generalize well to all datasets or domains. Another limitation is that while Stacking effectively combines predictions, it may struggle with datasets that have high levels of noise or class overlap, potentially reducing its interpretability and robustness. A more explicit focus on these aspects in future work would provide a balanced view and further improve its practical applicability.

5. Conclusion

This study successfully addressed the challenges of text classification on imbalanced data by applying the SMOTE method to balance class distribution and evaluating the performance of various individual classification algorithms as well as ensemble methods. The results demonstrate that SMOTE effectively balances the number of samples in each class, allowing machine learning models to better learn patterns from all classes while reducing bias toward the majority class. Based on testing results, the Stacking algorithm delivered the best performance, achieving an accuracy, precision,

recall, and F1-score of 88%. This is due to its ability to combine the strengths of multiple base models (CART, KNN, SVM, RF) using Logistic Regression as the meta-classifier. Bagging with Random Forest ranked second, demonstrating stable performance with an accuracy of 86%. Meanwhile, Hard Voting produced fairly good results with an accuracy of 84%, although it is less flexible than stacking in handling data complexity. Soft Voting had the lowest performance, with an accuracy of 81%, indicating that this method is less optimal for complex datasets. Additionally, individual algorithms like MNB showed relatively good performance compared to KNN but remained below the performance of ensemble methods. Overall, this study proves that the application of SMOTE and ensemble techniques, particularly Stacking, can significantly enhance accuracy and prediction quality in text classification tasks.

However, this study has certain limitations. First, the ensemble methods, particularly Stacking, require significant computational resources and longer training times, which may hinder their scalability for larger datasets. Second, the reliance on traditional machine learning algorithms limits the ability to capture more complex patterns in textual data. Lastly, the generalizability of the models was not extensively tested across datasets from different domains, which may impact their robustness. Future research should address these limitations by exploring more efficient ensemble techniques, such as lightweight stacking frameworks or distributed training approaches, to reduce computational overhead. Additionally, integrating advanced deep learning models, such as LSTM or transformer-based architectures like BERT, could provide better context understanding and performance. Testing the models on diverse datasets and applying domain adaptation techniques would further enhance their applicability and robustness.

6. Declarations

6.1. Author Contributions

Conceptualization: P.P.P., M.K.A., A.S.C., A.H., N.H., A.M.; Methodology: A.M.; Software: P.P.P.; Validation: P.P.P., A.M., and N.H.; Formal Analysis: P.P.P., A.M., and N.H.; Investigation: P.P.P.; Resources: A.M.; Data Curation: A.M.; Writing Original Draft Preparation: P.P.P., A.M., and N.H.; Writing Review and Editing: A.M., P.P.P., and N.H.; Visualization: P.P.P.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Achmad and I. I. Wiratmadja, "Strategic advancements in tourism development in Indonesia: Assessing the impact of facilities and services using the PLS-SEM approach," *Journal Industrial Servicess*, vol. 10, no. 1, pp. 49–62, Apr. 2024, doi: 10.62870/jiss.v10i1.24494.

[2] A. Arumugam, S. Nakkeeran, and R. Subramaniam, "Exploring the Factors Influencing Heritage Tourism Development: A Model Development," *Sustainability (Switzerland)*, vol. 15, no. 15, pp. 1–18, Aug. 2023, doi: 10.3390/su151511986

[3] M. S. Antonio and M. M. Uula, "Measuring the Productivity of Tourism Sector in Indonesia," *Halal Tourism and Pilgrimage*, vol. 2, no. 2, pp. 1–11, Dec. 2022, doi: 10.58968/htp.v2i2.179.

[4] P. Candra Susanto, M. Rizky Mahaputra, and M. Ridho Mahaputra, "Service Quality and Customer Satisfaction Have an Impact on Increasing Hotel Room Occupancy Ratio: Literature Review Study," *Greenation International Journal of Tourism and Management*, vol. 1, no. 4, pp. 400–412, Jan. 2024, doi: 10.38035/gijtm.v1i4.127.

[5] H. J. Christanto and Y. A. Singgalen, "Sentiment Analysis of Customer Feedback Reviews Towards Hotel's Products and Services in Labuan Bajo," *Journal of Information Systems and Informatics*, vol. 4, no. 4, pp. 805–822, Dec. 2022, doi: 10.51519/journalisi.v4i4.294.

[6] K. Puh and M. Bagić Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *Journal of Hospitality and Tourism Insights*, vol. 6, no. 3, pp. 1188–1204, Jun. 2023, doi: 10.1108/JHTI-02-2022-0078.

[7] P. Rita, R. Ramos, M. T. Borges-Tiago, and D. Rodrigues, "Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study," *Int J Hosp Manag*, vol. 104, no. 1, pp. 1–12, Jul. 2022, doi: 10.1016/j.ijhm.2022.103245.

[8] E. Sihombing, M. Halmi Dar, and F. A. Nasution, "Comparison Of Machine Learning Algorithms In Public Sentiment Analysis Of TAPERA Policy," *International Journal of Science, Technology & Management*, vol. 5, no. 5, pp. 1089–1098, Sep. 2024, doi: 10.46729/ijstm.v5i5.1164.

[9] R. Kusumaningrum, I. Z. Nisa, R. P. Nawangsari, and A. Wibowo, "Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, pp. 292–303, Nov. 2021, doi: 10.26555/ijain.v7i3.737.

[10] N. Novia Hidayati, "Aspect-Based Sentiment Analysis for Hotel Reviews with Latent Dirichlet Allocation and Machine Learning Algorithms," *Register*, vol. 9, no. 2, pp. 144–159, Jul. 2023, doi: 10.26594/register.v9n2.3441.

[11] M. K. Anam, M. B. Firdaus, F. Suandi, Lathifah, T. Nasution, and S. Fadly, "Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining," in *ICFTSS 2024 - International Conference on Future Technologies for Smart Society*, Kuala Lumpur: Institute of Electrical and Electronics Engineers Inc., vol. 2024, no. Sep., pp. 90—95, Sep. 2024. doi: 10.1109/ICFTSS61109.2024.10691363.

[12] M. K. Anam, S. Defit, Haviluddin, L. Efrizoni, and M. B. Firdaus, "Early Stopping on CNN-LSTM Development to Improve Classification Performance," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1175–1188, 2024, doi: 10.47738/jads.v5i3.312.

[13] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, no. 1, pp. 1–30, Mar. 2024, doi: 10.1016/j.nlp.2024.100059.

[14] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach Learn*, vol. 113, no. 1, pp. 4845–4901, Jul. 2022, doi: 10.1007/s10994-022-06268-8.

[15] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1272–1285, Sep. 2024, doi: 10.47738/jads.v5i3.316.

[16] M. B. Ressan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 375–383, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[17] N. W. Susanto and H. Suparwito, "SVM-PSO Algorithm for Tweet Sentiment Analysis #BesokSenin," *Indonesian Journal of Information Systems (IJIS)*, vol. 6, no. 1, pp. 36–47, 2023, doi: 10.24002/ijis.v6i1.7551.

[18] F. P. Arifianti and A. Salam, "XGBoost and Random Forest Optimization using SMOTE to Classify Air Quality," *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, pp. 1–8, Nov. 2024, doi: 10.26877/asset.v6i1.18136.

[19] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, "Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, Aug. 2023, doi: 10.51519/journalisi.v5i3.539.

[20] Y. Q. Song, X. Yao, Z. Liu, X. Shen, and J. Mao, "An Improved C4.5 Algorthm in Bagging Integration Model," *IEEE Access*, vol. 8, no. 1, pp. 206866–206875, 2020, doi: 10.1109/ACCESS.2020.3032291.

[21] B. L. V. S. R. Krishna, V. Mahalakshmi, and G. K. M. Nukala, "A Stacking Model for Outlier Prediction using Learning Approaches," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 2s, pp. 629–638, 2023.

[22] S. Hadhri, M. Hadiji, and W. Labidi, "A voting ensemble classifier for stress detection," *Journal of Information and Telecommunication*, vol. 8, no. 3, pp. 1–18, 2024, doi: 10.1080/24751839.2024.2306786.

[23] R. Budiyanto, I. Purnamasari, and D. D. Saputra, "Text Mining for Customer Sentiment Using Naive Bayes and SMOTE Methods on TokopediaCare Twitter," *International Journal of Information System & Technology Akreditasi*, vol. 6, no. 1, pp. 134–144, 2022, doi: 10.30645/ijistech.v6i1.221.

[24] Y. A. Mustofa and I. S. K. Idris, "Ensemble Approach to Sentiment Analysis of Google Play Store App Reviews," *Journal of Electrical and Electronics Engineering*, vol. 6, no. 2, pp. 181–188, 2024, doi: 10.37905/jjeee.v6i2.25184.

[25] A. R. Susanti and E. N. Ilahi, "Sentiment Analysis of User Reviews of E-commerce Applications: Case Study on the Shoppe Platform," *Journal of Social Science*, vol. 5, no. 4, pp. 983–988, 2024, doi: 10.46799/jss.v5i4.885.

[26] N. S. Sediatmoko, Y. Nataliani, and I. Suryady, "Sentiment Analysis of Customer Review Using Classification Algorithms and SMOTE for Handling Imbalanced Class," *Indonesian Journal of Information Systems*, vol. 7, no. 1, pp. 38–52, 2024, doi: 10.24002/ijis.v7i1.8879.

[27] A. I. Gufroni, I. Hoeronis, N. Fajar, A. Rachman, C. M. S. Ramdani, and H. Sulastri, "Implementation of Ensemble Machine Learning Classifier and Synthetic Minority Oversampling Technique for Sentiment Analysis of Sustainable Development Goals in Indonesia," *International Journal on Informatics Visualization*, vol. 8, no. 2, pp. 678–685, 2024, doi: 10.62527/joiv.8.2.1949.

[28] A. K. Abbas, A. K. Salih, H. A. Hussein, Q. M. Hussein, and S. A. Abdulwahhab, "Twitter Sentiment Analysis Using an Ensemble Majority Vote Classifier," *Journal of Southwest Jiaotong University*, vol. 55, no. 1, 2020, doi: 10.35741/issn.0258-2724.55.1.9.

[29] M. K. Anam, L. L. Van FC, H. Hamdani, R. Rahmaddeni, J. Junadhi, M.B. Firdaus, I. Syahputra, Y. Irawan, "Sara Detection on Social Media Using Deep Learning Algorithm Development," *Journal of Applied Engineering and Technological Science*, vol. 6, no. 1, pp. 225–237, Dec. 2024, doi: 10.37385/jaets.v6i1.5390.

[30] M. K. Anam, Munawir, L. Efrizoni, N. Fadillah, W. Agustin, I. Syahputra, T. P. Lestari, M. B. Firdaus, Lathifah, A. K. Sari "Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 354–365, Jan. 2025, doi: 10.47738/jads.v6i1.459.

[31] K. Nurfebia and S. Sriani, "Sentiment Analysis of Skincare Products Using the Naive Bayes Method," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 1663–1676, Sep. 2024, doi: 10.51519/journalisi.v6i3.817.

[32] A. Majid, D. Nugraha, and F. D. Adhinata, "Sentiment Analysis on Tiktok Application Reviews Using Natural Language Processing Approach," *Journal of Embedded System Security and Intelligent System*, vol. 4, no. 1, pp. 32–38, 2023, doi: 10.26858/jessi.v4i1.41897.

[33] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesti, and A. Arifianto, "Sentiment Analysis on a Large Indonesian Product Review Dataset," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 167–178, 2024, doi: 10.20473/jisebi.10.1.167-178.

[34] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes," *Scientific Journal of Informatics*, vol. 10, no. 1, pp. 25–34, Jan. 2023, doi: 10.15294/sji.v10i1.39952.

[35] V. Talasila, M. V Mohan, and N. M. R, "Enhancing Text-to-Image Synthesis with an Improved Semi-Supervised Image Generation Model Incorporating N-Gram, Enhanced TF-IDF, and BOW Techniques," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 381–397, 2023

[36] M. Ibrahim, "Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective," *Dhaka University Journal of Applied Science and Engineering*, vol. 7, no. 1, pp. 66–71, Feb. 2023, doi: 10.3329/dujase.v7i1.62888.

[37] X. T. Dang and T. T. Le, "KNN-SMOTE: An Innovative Resampling Technique Enhancing the Efficacy of Imbalanced Biomedical Classification," in *Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications*, N. Hoang Phuong, N. T. Huyen Chau, and V. Kreinovich, Eds., Cham: Springer Nature Switzerland, pp. 111–121, 2024. doi: 10.1007/978-3-031-63929-6_11.

[38] M. Sakhdiah, A. Salma, D. Permana, and D. Fitria, "Sentiment Analysis Using Support Vector Machine (SVM) of ChatGPT Application Users in Play Store," *UNP Journal of Statistics and Data Science*, vol. 2, no. 2, pp. 151–158, May 2024, doi: 10.24036/ujsds/vol2-iss2/158

[39] L. L. Van FC, M. K. Anam, S. Bukhori, A. K. Mahamad, S. Saon, and R. L. V. Nyoto, "The Development of Stacking Techniques in Machine Learning for Breast Cancer Detection," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 71–85, Jan. 2025, doi: 10.47738/jads.v6i1.416.

[40] M. Atif, F. Anwer, and F. Talib, "An Ensemble Learning Approach for Effective Prediction of Diabetes Mellitus Using Hard Voting Classifier," *Indian J Sci Technol*, vol. 15, no. 39, pp. 1978–1986, 2022, doi: 10.17485/IJST/v15i39.1520.

[41] H. Ghali Jabbar, "Advanced Threat Detection Using Soft and Hard Voting Techniques in Ensemble Learning," *Journal of Robotics and Control (JRC)*, vol. 5, no. 4, pp. 1104–1116, 2024, doi: 10.18196/jrc.v5i4.22005.

[42] S. G. Begum and P. K. Sree, "Drug Recommendation Using a 'Reviews and Sentiment Analysis' By a Recurrent Neural Network," *Indonesian Journal of Multidisciplinary Science*, vol. 2, no. 9, pp. 3085–3094, Jun. 2023, doi: 10.55324/ijoms.v2i9.530.

[43] M. Hung, E. Lauren, E. S. Hon, W. C. Birmingham, J. Xu, S. Su, S. D. Hon, J. Park, P. Dang, M. S. Lipsky, "Social network analysis of COVID-19 sentiments: Application of artificial intelligence," *J Med Internet Res*, vol. 22, no. 8, pp. 1–13, 2020, doi: 10.2196/22590.

[44] L. Zhu, M. Xu, Y. Bao, Y. Xu, and X. Kong, "Deep learning for aspect-based sentiment analysis: a review," *PeerJ Comput Sci*, vol. 8, no. 1, pp. 1–37, 2022, doi: 10.7717/PEERJ-CS.1044.

[45] Y. H. Hsieh and X. P. Zeng, "Sentiment Analysis: An ERNIE-BiLSTM Approach to Bullet Screen Comments," *Sensors*, vol. 22, no. 14, pp. 1–15, Jul. 2022, doi: 10.3390/s22145223.

[46] D. Hardiansyah, R. A. Aziz, and M. S. Hasibuan, "The Classification Method is Used for Sentiment Analysis in My Telkomsel," *International Journal of Artificial Intelligence Research*, vol. 8, no. 2, pp. 1–11, Dec. 2024, doi: 10.29099/ijair.v8i2.1229.

[47] S. R. Aisy and B. Prasetiyo, "Sentiment Analysist of the TPKS Law on Twitter Using InSet Lexicon with Multinomial Naïve Bayes and Support Vector Machine Based on Soft Voting," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 93–101, Sep. 2023, doi: 10.15294/rji.v1i2.68324.

[48] R. Jayapermana, A. Aradea, and N. I. Kurniati, "Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topics on Twitter," *Scientific Journal of Informatics*, vol. 9, no. 1, pp. 8–15, May 2022, doi: 10.15294/sji.v9i1.31648.

[49] I. R. Munthe, B. H. Rambe, F. Hanum, A. T. Amanda, A. S. R. Hutagaol, and R. Harianto, "Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 2079–2091, Dec. 2024, doi: 10.47738/jads.v5i4.421.