



Evaluating Machine Learning Approaches for Sentiment Analysis of Internet Service Providers in Indonesia: Naïve Bayes vs. Gradient Boosting

I Gusti Ngurah Bagus Catur Bawa^{1*}, Made Pradnyana Ambara²,
I Wayan Suasnawa³, Anak Agung Ngurah Gde Sapteka⁴,
I Komang Wiratama⁵, and Ida Bagus Putra Manuaba⁶

^{1,2,3,5,6}Information Technology Department, Politeknik Negeri Bali, Bali, Indonesia

⁴Electrical Engineering Department, Politeknik Negeri Bali, Bali, Indonesia
caturbawa@pnb.ac.id

Abstract. Customer satisfaction has become a crucial metric for the business sustainability of Internet Service Providers (ISPs) amidst fierce industrial competition in Indonesia. Automated sentiment analysis on social media offers a solution to efficiently monitor public opinion. This study aims to comprehensively evaluate and compare two machine learning algorithms, Naïve Bayes (NB) and Gradient Boosting (GB), for the task of sentiment classification on Indonesian ISP user reviews collected from the Twitter platform. The research methodology encompasses several optimization stages, including hyperparameter tuning and the handling of imbalanced data using the Synthetic Minority Over-sampling Technique (SMOTE), to ensure a fair and in-depth comparison. The experimental results indicate that Gradient Boosting consistently outperforms Naïve Bayes. The best-performing model, Gradient Boosting optimized with tuning and SMOTE, achieved an overall accuracy of 85%. More importantly, this model demonstrated a superior capability in identifying negative sentiment, achieving a recall score of 93%, which is a valuable capability for practical applications such as customer complaint detection. This study concludes that the optimized Gradient Boosting approach constitutes a more robust and reliable solution for sentiment analysis in this domain.

Keywords: Gradient Boosting, Naïve Bayes Sentiment Analysis, Machine Learning

1 Introduction

The rapid advancement of digital technology in Indonesia has solidified the role of Internet Service Providers (ISPs) as critical infrastructure. In an increasingly competitive market, customer satisfaction has become a determining factor for maintaining long-term business sustainability. Consequently, tracking and interpreting public sentiment is no longer a secondary concern but a strategic imperative. In this context, social media platforms such as Twitter (X) have emerged as key channels for

© The Author(s) 2025

A. A. N. G. Sapteka et al. (eds.), *Proceedings of the International Conference on Sustainable Green Tourism Applied Science - Engineering Applied Science 2025 (ICOSTAS-EAS 2025)*, Advances in Engineering Research 280,

https://doi.org/10.2991/978-94-6463-878-3_58

capturing real-time and unsolicited feedback from users. However, the immense volume of user-generated content, which primarily consists of unstructured text, presents a substantial obstacle to traditional manual analysis techniques (Kemp, 2024). To overcome this limitation, automated sentiment analysis—particularly those leveraging Natural Language Processing (NLP)—has been proposed as an effective alternative. These techniques facilitate large-scale opinion classification and can provide actionable insights to support decision-making and enhance customer engagement strategies (Liu, 2012).

Various machine learning algorithms have been utilized for sentiment analysis tasks. In the Indonesian setting, earlier studies have often compared conventional models such as Naïve Bayes (NB), known for its simplicity and computational efficiency (Rish, 2001), with more complex alternatives like Support Vector Machine (SVM). Several findings have indicated that SVM may offer a marginal performance advantage when applied to sentiment analysis involving ISP-related data (Pamungkas et al., 2021). In broader applications, ensemble-based approaches such as Gradient Boosting (GB) have consistently demonstrated high performance by integrating multiple weak learners into a stronger predictive model (Natekin & Knoll, 2013). Comparative research has shown that GB can outperform NB in many scenarios (Shachi & Kumar, 2022). Despite this, there remains a noticeable absence of comprehensive, head-to-head comparisons between NB and GB that specifically focus on Indonesian ISP-related sentiment datasets. This is especially true when accounting for the effects of optimization techniques. This gap is critical, considering that real-world social media data tends to be noisy, highly imbalanced, and often expressed in informal language—factors that significantly influence model performance (Go et al., 2009; Sun et al., 2009).

To address this gap, the current study conducts a detailed evaluation and comparison of the Naïve Bayes and Gradient Boosting algorithms in the context of sentiment classification for user comments about Indonesian ISPs. The analysis not only focuses on core performance metrics but also considers the impact of hyperparameter tuning and the application of the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance. The primary aim is to derive deeper insights into the relative strengths and limitations of each algorithm and, based on these findings, to provide an evidence-based recommendation for real-world implementation in Indonesia's telecommunications sector.

2 Methodology

This research commenced with a data acquisition phase that focused on gathering user-generated content related to the services provided by major Internet Service Providers (ISPs) in Indonesia. The dataset was sourced from Twitter (X), using the Python-based `snsrape` library to extract tweets containing specific keywords associated with ISP providers. The data collection spanned the period from January 1 to December 31, 2024. After eliminating duplicate entries, the dataset was prepared for the next stage of processing. Following this, a manual sentiment labeling process was carried out by the research team. Each tweet was categorized into one of three sentiment classes: Positive,

Negative, or Neutral. This annotation process resulted in a curated dataset comprising 2,967 labeled samples, which subsequently served as the input for the training and evaluation of machine learning models. It is worth noting that the class distribution within the dataset was imbalanced, a condition frequently observed in naturally occurring social media data.

2.1 Data Collection and Labeling

The research began with the collection of data. The dataset comprised user-generated reviews about the services of leading Internet Service Providers (ISPs) in Indonesia, gathered from the social media platform Twitter (X). To extract the data, the study utilized the `snsrape` library in Python, targeting tweets that included specific ISP names as keywords, posted between January 1 to December 31, 2024. Following the elimination of duplicate entries, the remaining dataset formed the basis for annotation. Researchers then performed manual labeling, categorizing each tweet into one of three sentiment classes: Positive, Negative, or Neutral. This labeling process resulted in a final annotated dataset of 2,967 samples, which was then used in the experimental phase. As is common in real-world opinion data, the distribution of sentiment classes was imbalanced.

2.2 Preprocessing and Feature Extraction

The raw textual data collected for this study were processed through a conventional text preprocessing workflow designed to enhance data quality prior to feature extraction, as recommended by Koto and Rahman (2017). This procedure involved several sequential steps: converting all text to lowercase (case folding), removing extraneous elements such as user handles, hashtags, hyperlinks, and punctuation (cleansing), eliminating common Indonesian stopwords through a predefined stopword list, and applying stemming using the `Sastrawi` library to reduce words to their root forms (Tahitoe & Purwarianti, 2015). After these steps, the cleaned corpus was transformed into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This statistical approach assigns weight based on word importance across the corpus. To maintain computational efficiency and reduce feature sparsity, the dimensionality of the resulting feature space was restricted to the 5,000 most informative terms, following best practices as outlined by Manning et al. (2008).

2.3 Experimental Design and Modeling

The feature-vector dataset was split into training (80%) and testing (20%) sets using a stratified approach to maintain class proportions. The experimental design involved two primary optimization scenarios performed on the training data. First, hyperparameter tuning was conducted for both Naïve Bayes (`alpha`) and Gradient Boosting (`max_depth`, `n_estimators`) using `GridSearchCV` with 3-fold cross-validation. Second, the tuned models were retrained on a balanced dataset created by applying the SMOTE technique (Chawla et al., 2002) to the original training set. All models were implemented using the `Scikit-learn` library (Pedregosa et al., 2011).

2.4 Performance Evaluation

To objectively assess and compare the classifiers, this study employed four standard evaluation metrics frequently used in supervised machine learning (Fawcett, 2006). These include accuracy, which captures the percentage of correct predictions across all instances; precision, which evaluates the model's accuracy in predicting positive labels; recall, which measures the ability to retrieve all relevant positive examples; and F1-score, which balances precision and recall through a harmonic mean.

Each metric was calculated both per class and as a weighted average, allowing for a more nuanced assessment that considers class imbalances. In addition to these numerical indicators, confusion matrices were generated to provide a visual overview of classification outcomes and error distributions.

3 Result and Discussion

This section presents the empirical results obtained from the experimental procedures and offers a detailed interpretation of the findings. The primary analysis centers on evaluating and comparing the Naïve Bayes and Gradient Boosting algorithms under different tuning and sampling strategies. The discussion highlights the significance of the quantitative results, connects them with existing research, and explores the practical implications of selecting the most effective model.

3.1 Baseline Performance of Tuned Models

The initial experiment compared the performance of the Naïve Bayes and Gradient Boosting models following a hyperparameter optimization process. As summarized in Table 1, the results indicate a significant performance advantage for the Gradient Boosting model.

As summarized in Table 1, the Gradient Boosting model exhibits a notable edge in performance compared to its counterpart. After hyperparameter tuning, the Gradient Boosting classifier achieved an overall accuracy of 84%, surpassing the Naïve Bayes model, which attained 79%. This advantage is further evident in the weighted average F1-score, where Gradient Boosting reached 83%, while Naïve Bayes recorded 80%, suggesting that the former offers more consistent and balanced performance across sentiment classes. A closer inspection of class-level results reveals that Gradient Boosting is particularly effective in managing the neutral category, achieving an F1-score of 77%, substantially higher than the 70% recorded by Naïve Bayes. These observations are consistent with the findings reported by Shachi and Kumar (2022) and support broader insights into the superior capability of ensemble-based approaches when applied to nuanced textual data (Dietterich, 2000). Despite its impressive recall of 96% in identifying negative sentiments, which highlights Naïve Bayes' strength in detecting user complaints, its comparatively weaker performance in other categories diminishes its overall dependability.

Table 1. Performance Comparison of Tuned Models

Metric	Naïve bayes (Tuned)			Gradient boosting (Tuned)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Negative	0.77	0.96	0.86	0.84	0.81	0.82
Neutral	0.65	0.77	0.70	0.83	0.72	0.77
Positive	0.91	0.75	0.82	0.84	0.90	0.87
Accuracy			0.79			0.84
Weighted Avg	0.81	0.79	0.80	0.84	0.84	0.83

Table 2. Performance Comparison of Models after SMOTE Implementation

Metric	Naïve bayes (Tuned + SMOTE)			Gradient boosting (Tuned + SMOTE)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Negative	0.78	0.97	0.87	0.71	0.93	0.80
Neutral	0.64	0.75	0.69	0.83	0.77	0.80
Positive	0.90	0.75	0.82	0.92	0.86	0.89
Accuracy			0.79			0.85
Weighted Avg	0.81	0.79	0.79	0.86	0.85	0.85

Table 2 reveals that the most notable effect of the applied optimization was observed in the Gradient Boosting model. Its overall accuracy experienced a modest yet meaningful increase, reaching 85%. More importantly, the recall score for the negative sentiment class underwent a substantial improvement, rising from 81% to 93%. This leap is especially significant from a practical standpoint, as it indicates the model's enhanced capacity to detect the vast majority of user complaints correctly. Such a shift suggests a greater readiness for deployment in contexts where identifying dissatisfaction is critical. This gain in recall, however, came with a corresponding reduction in precision, illustrating the classic precision-recall trade-off discussed by Davis and Goadrich (2006). In practical applications, particularly within customer service frameworks, this trade-off is generally acceptable or even desirable, given that failing to detect genuine negative feedback (false negatives) can carry more serious consequences than investigating an incorrect alert (false positives). In contrast, the application of SMOTE to the Naïve Bayes model did not produce any notable increase in overall accuracy. This outcome suggests that, despite addressing class imbalance, the method was insufficient to overcome the inherent limitations of Naïve Bayes in handling complex sentiment patterns.

3.2 The Impact of Handling Class Imbalance with SMOTE

One of the most prominent issues encountered in the dataset was the imbalance among sentiment classes, with positive reviews significantly outnumbering both neutral and

negative ones. To mitigate this disparity, the Synthetic Minority Over-sampling Technique (SMOTE) was applied specifically to the training portion of the data. The effectiveness of this approach in achieving a more balanced distribution across sentiment categories is visually demonstrated in Figure 1, which highlights the post-processing class proportions.

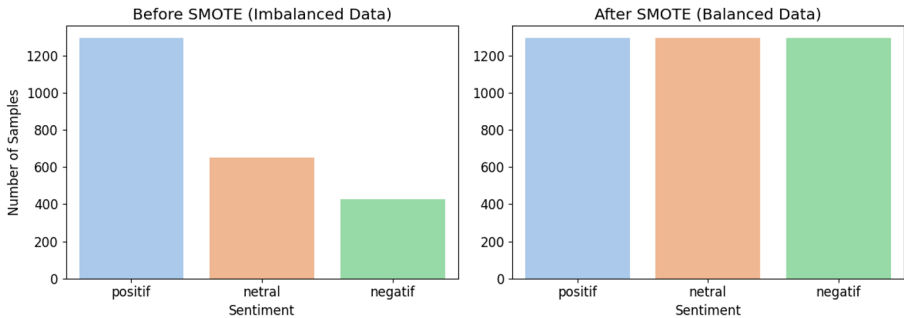


Figure 1. Class Distribution Before and After SMOTE

Training the models on the balanced dataset yielded a notable shift in performance, with the Gradient Boosting model benefiting most prominently from this adjustment. The outcomes of this training scenario are detailed in Table 2, which presents the evaluation results for both algorithms after being trained on the dataset that had been resampled using the SMOTE technique.

3.3 Comprehensive Evaluation and Feature Importance

As part of the final evaluation, Figure 2 provides a comparative visualization of the performance across all four model configurations.

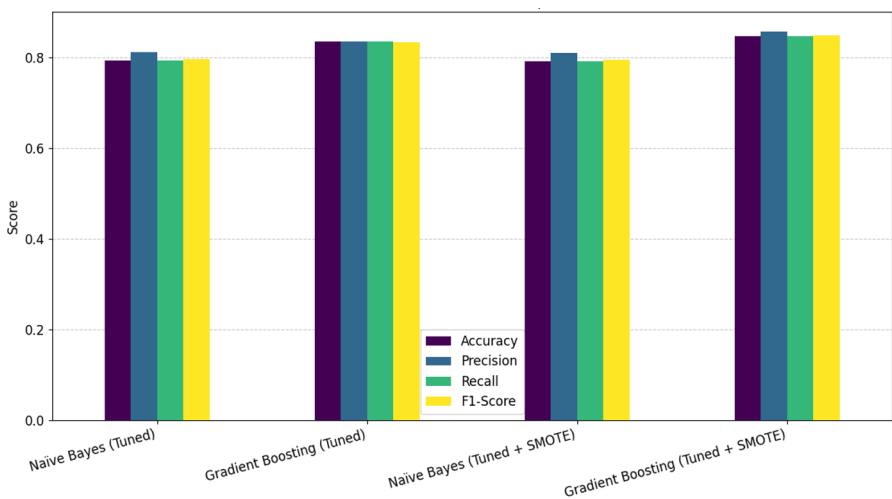


Figure 2. Model Performance Comparison Chart

This summary illustration reinforces the conclusion that the Gradient Boosting model, when optimized through hyperparameter tuning and enhanced with SMOTE, consistently outperforms the other configurations across multiple evaluation criteria.

4 Conclusion

This research aimed to conduct a systematic evaluation and comparison of the Naïve Bayes and Gradient Boosting algorithms in the context of sentiment classification for user feedback directed at Indonesian Internet Service Providers. The results clearly indicate that Gradient Boosting, particularly when enhanced through hyperparameter optimization and the application of SMOTE, substantially outperforms Naïve Bayes. While the model's 85% overall accuracy is noteworthy, its practical impact is even more significant. Specifically, its ability to correctly identify 93% of negative sentiment cases makes it a valuable asset for operational tasks such as customer complaint monitoring. This high recall rate, achieved through a strategic approach to class imbalance, reinforces the model's applicability in real-world business scenarios. Additionally, feature importance analysis validates that semantically meaningful and contextually appropriate terms inform the model's predictions.

Although the study achieves its intended goals, the findings also point to promising directions for further investigation. Future research could consider integrating more sophisticated deep learning models, such as IndoBERT, to better capture the subtleties of the Indonesian language (Wilie et al., 2020). In addition, transitioning from a document-level analysis to an Aspect-Based Sentiment Analysis (ABSA) framework may yield richer insights by identifying specific service aspects that influence user satisfaction (Pontiki et al., 2014).

References

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Davis, J., & Goadrich, M. (2006). The Relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. <https://doi.org/10.1145/1143844.1143874>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1–15. https://doi.org/10.1007/3-540-45014-9_1.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).
- Kemp, S. (2024). *Digital 2024: Indonesia*. Data Reportal. <https://datareportal.com/reports/digital-2024-indonesia>.

- Koto, F., & Rahman, G. (2017). The Effect of pre-processing in sentiment analysis for Indonesian text. *Procedia Computer Science*, 116, 559–565. <https://doi.org/10.1016/j.procs.2017.10.046>.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines. A tutorial. *Frontiers in Neuroinformatics*, 7(21). <https://doi.org/10.3389/fnbot.2013.00021>.
- Pamungkas, A. S., Permanasari, A. E., & Hidayatullah, A. F. (2021). Sentiment analysis on Twitter data for an Indonesian internet service provider using Naive Bayes and Support Vector Machine. *Journal of Physics: Conference Series*, 1844(1), 012023. <https://doi.org/10.1088/1742-6596/1844/1/012023>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect-based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.
- Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3, 41–46.
- Shachi, R. K., & Kumar, S. (2022). A comparative study of Naive Bayes, Support Vector Machine, and Gradient Boosting for sentiment analysis on movie reviews. *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 450–455. <https://doi.org/10.1109/COM-IT-CON54601.2022.9850774>.
- Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), 687–719. <https://doi.org/10.1142/S021800140900732X>.
- Tahitoe, D. A., & Purwarianti, A. (2015). The development of an Indonesian language stemmer based on suffixed derivational and inflectional morphology. *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 123–128.
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Ma, J., Fung, P., & Prescher, S. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

