The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 5, October 2024
https://doi.org/10.33022/ijcs.v13i5.4386

Comparative Analysis of Machine Learning Algorithms with SMOTE and Boosting Techniques in Accuracy Improvement

Yuda Irawan[1*], Refni Wahyuni[2], Rian Ordila[3], Herianto[4]
yudairawan89@gmail.com[1*], refniabid@gmail.com[2], rian.68x@gmail.com[3], herianto.sy@gmail.com[4]
[1,2]Dept of Computer Science, Ilmu Komputer, Universitas Hang Tuah Pekanbaru
[3,4]Dept of Information System, Ilmu Komputer, Universitas Hang Tuah Pekanbaru

Article Information

Received : 12 Aug 2024
Revised : 9 Sep 2024
Accepted : 3 Oct 2024

Keywords

Random Forest, Naïve Bayes, SMOTE, XGBoost, Machine Learning

Abstract

This research explores and enhances accuracy in sentiment classification related to Indonesia's Capital City relocation by combining Naive Bayes (NB), Random Forest (RF), SMOTE, and XGBoost. The study addresses challenges of unbalanced data and complexity in social media sentiment analysis. The combination of RF with SMOTE achieved the highest accuracy at 91.25%, demonstrating SMOTE's effectiveness in balancing the dataset and improving minority class detection. While adding XGBoost slightly reduced accuracy (90.92%), it increased the NB model's accuracy from 77.45% to 85.97% when combined with SMOTE. RF alone reached 87.46% and improved to 88.78% with XGBoost. The study underscores the importance of selecting and combining techniques to maximize sentiment prediction accuracy. Future research could explore deep learning or transformer models for even better results, offering deeper insights into public sentiment and aiding effective policy strategy development.

A. Introduction

The development of information and communication technology has changed the way we collect, process, and analyze data. In this context, machine learning (ML) plays an important role in data-driven decision making. ML algorithms such as Naive Bayes and Random Forest are often used due to their ability to process complex and diverse data[1][2]. However, the performance of these algorithms can be compromised by the problems of data imbalance and poor data quality[3].

To overcome the data imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) has been developed[4]. SMOTE improves classification performance by adding synthetic samples to the minority class so that the data distribution becomes more balanced[5]. In addition, boosting techniques such as XGBoost have shown significant performance improvements in various machine learning applications by improving prediction accuracy through the combination of multiple weak models[6].

This research will conduct a comparative analysis between Naive Bayes and Random Forest algorithms with the application of SMOTE and Boosting techniques using XGBoost. The dataset used is data from Twitter regarding Sentiment Analysis of the Relocation of the National Capital City (IKN) in Indonesia. The relocation of IKN is a topic that is often discussed on social media and can provide insight into public sentiment regarding the policy[7].

This research is important to identify the most effective algorithms and techniques in processing imbalanced and complex data, and to provide recommendations that can be implemented in various data analysis contexts in Indonesia. In previous studies, the Random Forest algorithm has often performed better than other algorithms in handling imbalanced data, especially when combined with the SMOTE technique. However, with the increasing use of boosting techniques, such as XGBoost, further evaluation is needed to determine which algorithm gives the best results in specific contexts.

Thus, this research aims to make a significant contribution to the field of data science, particularly in the comparative analysis of machine learning algorithms with the application of SMOTE and Boosting techniques, as well as to provide a deeper insight into public sentiment towards important issues in Indonesia.

This research offers a significant contribution by combining two key machine learning techniques, namely SMOTE and Boosting (XGBoost), in the context of sentiment analysis on Twitter datasets related to important issues in Indonesia. In the last five years, various studies have examined the effectiveness of SMOTE in handling data imbalance. For example, research by Khariul Anam [8] showed a significant increase in the accuracy of prediction models when using SMOTE compared to the use of machine learning algorithms without SMOTE[9]. In addition, Ahmad Taufiq [10] emphasized the importance of SMOTE in improving classification performance on imbalanced medical data.

Other studies have shown that boosting techniques, especially XGBoost, can improve prediction accuracy in various machine learning applications[11]. Another study showed that XGBoost consistently outperformed other algorithms in a data mining competition[12]. Further researchers have also found that XGBoost provides the best results in credit risk prediction compared to algorithms

such as Random Forest[13]. The application of XGBoost to sentiment analysis showed a significant increase in accuracy compared to traditional models[14].

This research is unique because it combines both SMOTE and XGBoost techniques in sentiment analysis on datasets from social media that are relevant to public policy issues in Indonesia such as the relocation of the Ibu Kota Nusantara (IKN). By combining these two techniques Naïve Bayes algorithm and Random Forest this research is expected to improve the accuracy in classification and get a comparison of the best model usage.

B. Research Method

This research method is designed to improve accuracy and reliability in sentiment analysis using Twitter data. The research phase begins with the collection of relevant datasets, followed by a thorough text preprocessing process to clean and normalize the data. After that, the data was labeled using Lexicon Vader and weighted using TF-IDF technique. The dataset was then divided into training and testing data at a ratio of 70:30, to ensure a fair evaluation of the developed model. Two main algorithms, Naive Bayes and Random Forest, were applied with the addition of SMOTE technique to overcome data imbalance and XGBoost to improve prediction accuracy. Model performance results are compared to determine the best model, with the ultimate goal of producing an optimal sentiment analysis model. The following is a picture of the stages of research that will be carried out:

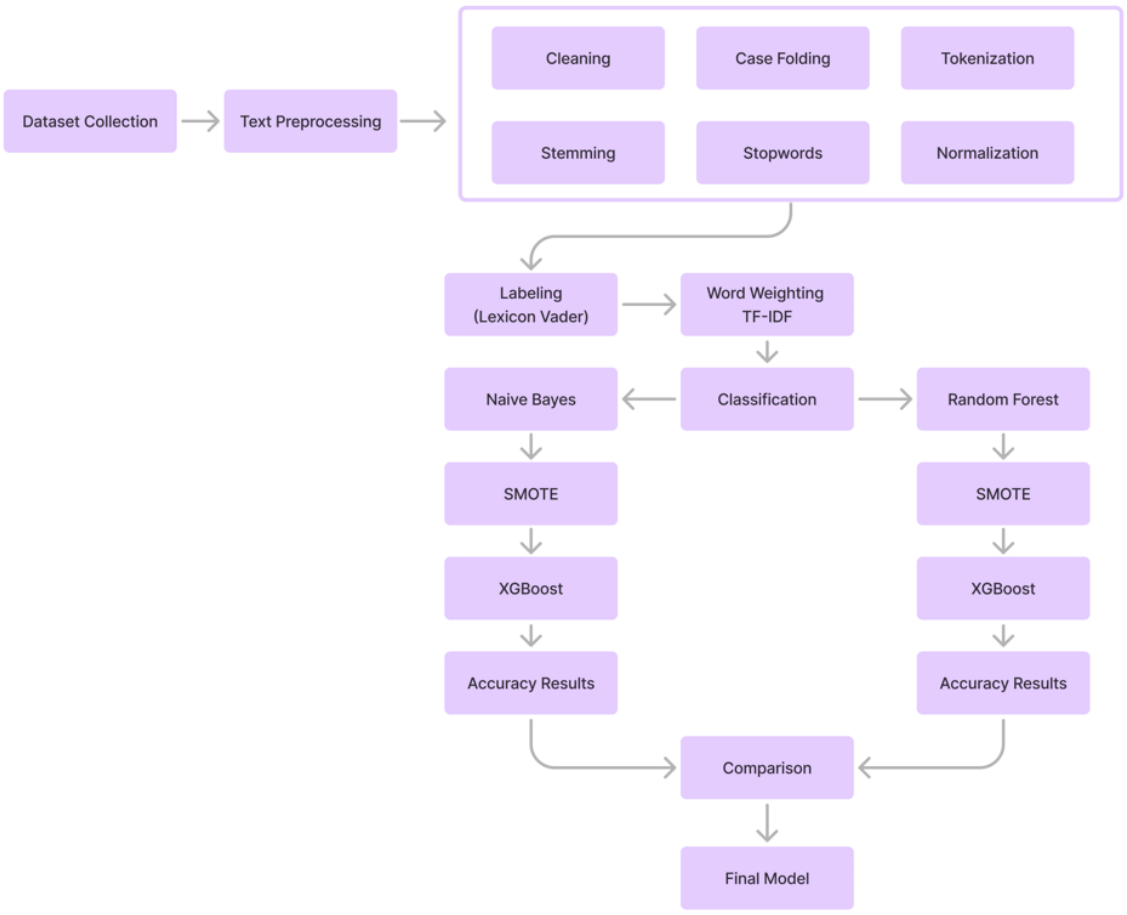

Figure 1. Research Methods

The research method shown in the figure includes various important stages in the data analysis process using machine learning techniques. The following is a detailed explanation of each stage:

1. Dataset Collection
Datasets were collected from Twitter containing public sentiments related to the relocation of the IKN.

2. Text Preprocessing
Text Preprocessing is the process of transforming, analyzing, and manipulating text to produce useful information. This process includes various techniques and methods used to process text, be it for data analysis, natural language processing (NLP), or other applications. Here are some of the steps in text processing:

- Cleaning: Cleaning the data from unnecessary characters or symbols to ensure that the data used is relevant and clean from noise[15].
- Case Folding: Changing all text to lowercase for consistency, so that 'Indonesia' and 'indonesia' are considered the same[16]
- Tokenization: Breaking text into small units (tokens) such as words or phrases to facilitate analysis[17].
- Stemming: Converting words to their stem form so that words like 'ran' and 'ran' are considered the same.
- Stopwords Removal: Removing common words that do not have important meaning in the analysis (e.g. 'and', 'in') to focus on more significant words[18].
- Normalization: Standardizing text formatting, such as converting numbers to a standardized form, for consistency and ease of further analysis[19].

3. Labeling
Using Lexicon Based to label the sentiment of the text (positive, negative, neutral) [20]. This method utilizes a predefined dictionary of words to assess the sentiment polarity of the text. Lexicon Based assigns scores based on the words found in the text and determines the overall sentiment based on the accumulation of those scores.

4. Word Weighting (TF-IDF)
Calculates the Term Frequency-Inverse Document Frequency (TF-IDF) value for each word in the document to determine the importance of the word in the document[21]. TF-IDF helps identify unique and relevant words in each document.

5. Data Splitting
The dataset is divided into two parts with a ratio of 70:30, where 70% of the data is used for model training and 30% for model testing. This split ensures that the model is trained with enough data and tested with representative data[22].

6. Classification

Naive Bayes: The first model applied is Naive Bayes for sentiment classification. Naive Bayes is a probabilistic-based classification algorithm that uses Bayes' Theorem with a strong assumption of independence between features. Although this independence assumption is rarely met in the real world, Naive Bayes often gives excellent results in various applications, such as text classification and sentiment analysis. The algorithm calculates the posterior probability of each class based on the given feature values, and then selects the class with the highest probability as the prediction. The advantages of Naive Bayes lie in its simplicity, computational efficiency, and ability to handle datasets with a large number of features[23].

Random Forest: The second model applied is Random Forest for sentiment classification. Random Forest is an ensemble learning method used for classification and regression. The algorithm works by building many decision trees during the training phase and the outputs of the individual classes of the trees are combined (averaged or majority voted) to determine the final prediction. Random Forest reduces the overfitting that often occurs with a single decision tree by combining predictions from multiple trees trained with different subsets of data and randomly selected features. This technique improves prediction accuracy and is resistant to overfitting on large and complex datasets[24].

7. Application of SMOTE
Synthetic Minority Over-Sampling Technique (SMOTE) is applied to handle data imbalance by adding synthetic samples to minority classes[25]. This technique helps in reducing the bias that may occur due to unbalanced sample size.

8. Application of XGBoost
Extreme Gradient Boosting (XGBoost) is used to improve model performance by combining several weak models to create a strong model[26]. XGBoost is known for its high speed and performance in various machine learning competitions.

9. Model Performance Evaluation
Each model is evaluated using metrics such as accuracy, precision, recall, and F1-score. This evaluation is important to measure the performance of the model and ensure that the selected model provides optimal results.

10. Comparison
The accuracy results and other metrics of the two models are compared to determine the best model. This comparison helps in selecting the most suitable model for the sentiment classification task on the dataset used.

11. Final Model
Based on the comparison results the best model is selected for final implementation and recommendation. This model can then be used for further analysis or implementation in real systems for sentiment classification.

C. Result and Discussion

This section explains the results and discussions that aim to answer all the questions in this study, namely the classification of public sentiment related to the relocation of the IKN in Indonesia using the Naive Bayes and Random Forest algorithms on Twitter social media.

Preprocessing and Labeling Stage

At this stage, the data obtained from Twitter undergoes a preprocessing process to clean the text from noise and eliminate irrelevant elements. The preprocessing stage begins with the collection of data from Twitter relating to public sentiment towards the relocation of the Capital City of the Archipelago (IKN) in Indonesia. The data obtained then underwent various cleaning stages to ensure the accuracy and relevance of the analysis. This process involves cleaning the text from special characters, symbols, numbers, as well as removing links and other irrelevant elements. Next, a case folding process is performed to convert the entire text into lowercase letters to ensure consistency. Tokenization is applied to break the text into units of words that are easier to analyze, while stopword removal is performed to eliminate common words that have no significant meaning in the analysis. Finally, stemming is applied to convert words into their base form, thus helping in simplifying the text and making it easier for the algorithm to identify patterns.

Labeling is done using a Lexicon-Based approach to identify the sentiment of each text into positive, negative, or neutral categories. Visualization of the sentiment distribution shows an imbalance with positive sentiment being more dominant than negative and neutral sentiment.

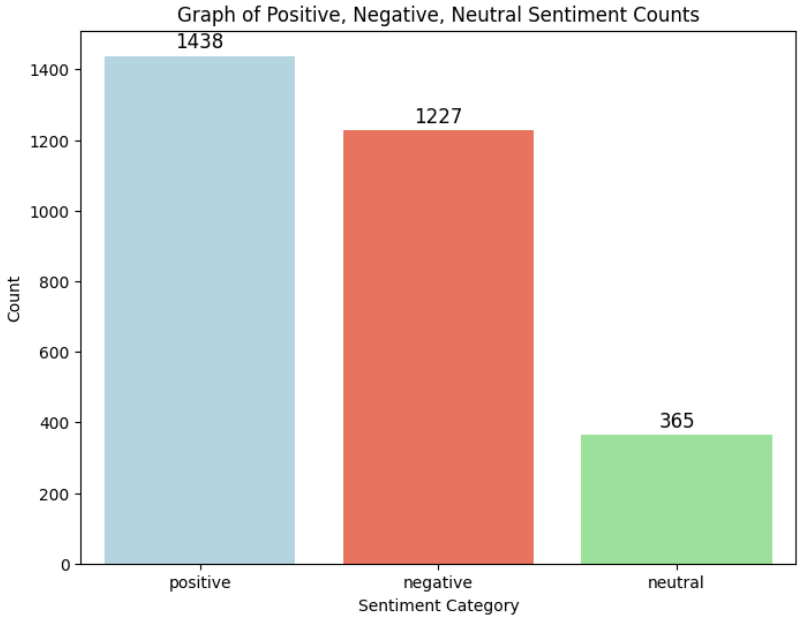

Figure 2. Sentiment Distribution Chart

The sentiment distribution chart shows that positive sentiments dominate Twitter discussions related to the relocation of Indonesia's capital city, with a total of 1,438, indicating the public's support or optimistic view of the policy. Negative sentiment came in second with 1,227, indicating concerns or criticism of the impacts of the move, such as environmental issues and costs. Neutral sentiment, totaling 365, reflects those with no strong or definite views. Overall, this graph illustrates the dominance of positive views although there remains significant criticism from social media users.

Model Testing and Evaluation

In the model testing stage, this research focuses on evaluating the performance of the Naive Bayes and Random Forest algorithms in sentiment analysis of Twitter data related to the relocation of the IKN. In addition, data imbalance handling techniques such as Synthetic Minority Over-Sampling Technique (SMOTE) and boosting algorithms such as Adaptive Boosting (AdaBoost) are also applied to improve prediction accuracy. This test aims to identify the best combination of algorithms and techniques to handle complex and imbalanced data, and provide further insight into the effectiveness of each approach in improving the performance of classification models.

Naïve Bayes and Random Forest Algorithms

Testing the Naive Bayes algorithm in this study was conducted to evaluate the model's ability to classify public sentiment related to the relocation of the IKN in Indonesia. Naive Bayes and Random Forest, which are known for their simple yet effective approach to text classification, were applied to understand how these models perform when faced with datasets containing positive, negative and neutral sentiments. This test focuses on the baseline accuracy of the algorithms before and after the application of performance enhancement techniques such as SMOTE and Boosting, and aims to identify potential improvements that can be achieved through the combination of these methods.

The confusion matrix analysis of the Naive Bayes and Random Forest models is shown in the following figure:

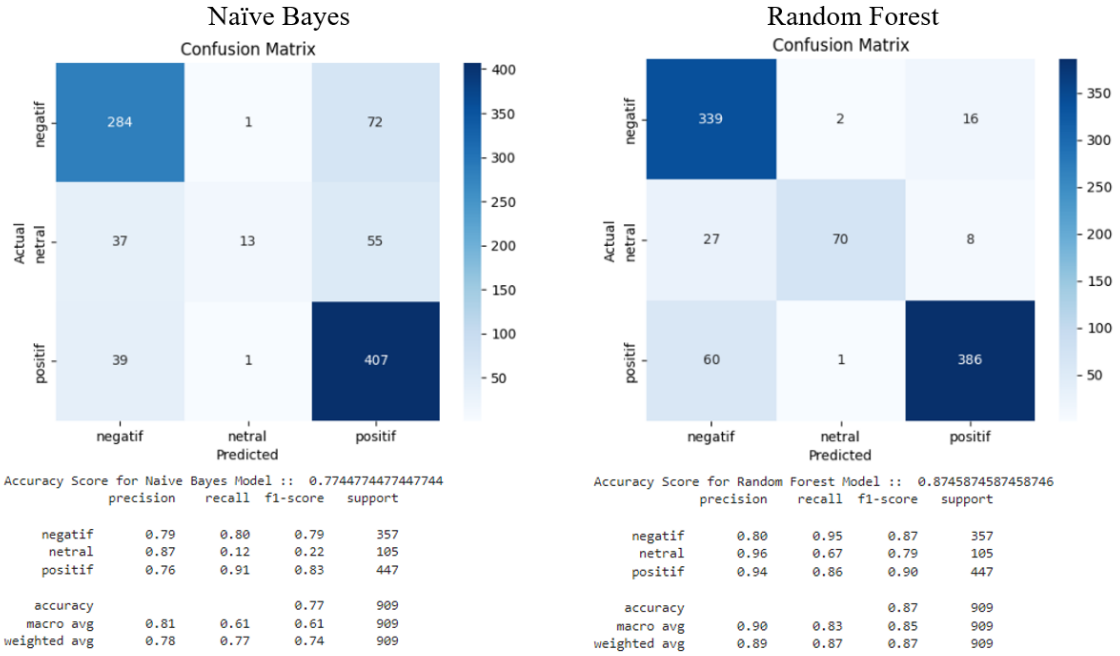

Figure 3. Comparison of Confusion Matrix 2 Algorithms

Confusion matrix and performance metrics for the Naive Bayes model show that it has an accuracy of 77.4%. Out of 909 samples, the model successfully classified 284 negative sentiments, 13 neutral sentiments, and 407 positive sentiments correctly. However, there were some misclassifications where 72 negative samples were misclassified as positive, and only 13 out of 105 neutral samples were correctly classified, showing the model's weakness in handling the neutral class. The highest precision was obtained in the neutral class (0.87), but with low recall (0.12), indicating that the model was very selective in predicting the neutral class but often missed many samples from the class. The highest F1-score was achieved in the positive class (0.83), indicating a better balance between precision and recall than the other classes. Overall, although the model has a fairly good accuracy, its performance can be improved, especially in dealing with neutral classes that are underrepresented in the prediction.

Based on the confusion matrix and performance metrics for the Random Forest model, the model shows excellent performance in the classification of sentiments related to the relocation of the Capital City of the Archipelago (IKN), with an accuracy of 87.5%. The model successfully classified 339 negative sentiments, 70 neutral sentiments, and 386 positive sentiments correctly from a total of 909 samples. The highest precision was obtained in the neutral class (0.96), indicating that the model was very accurate in predicting this class, although the recall for the neutral class was still low (0.67), indicating that there were still missed samples. The highest F1-score was achieved for the positive class (0.90), indicating a good balance between precision and recall for this class. The weighted average metric showed that the model as a whole provided excellent predictive results with an f1-score of 0.89. With high levels of accuracy and precision, the Random Forest model was able to effectively handle class imbalance, providing reliable and accurate predictions on this complex sentiment dataset.

Algorithms and XGBoost

This test will discuss the performance of the Naive Bayes and Random Forest algorithms combined with boosting techniques using XGBoost in the classification of public sentiment related to the relocation of the National Capital City (IKN) in Indonesia. This combination is expected to improve the accuracy and ability of the model to handle data imbalance better. The evaluation of the model's performance will be presented using a confusion matrix, which provides a detailed overview of the model's ability to accurately predict sentiment classes and shows the improvement resulting from the use of XGBoost compared to the basic Naive Bayes and Random Forest models.

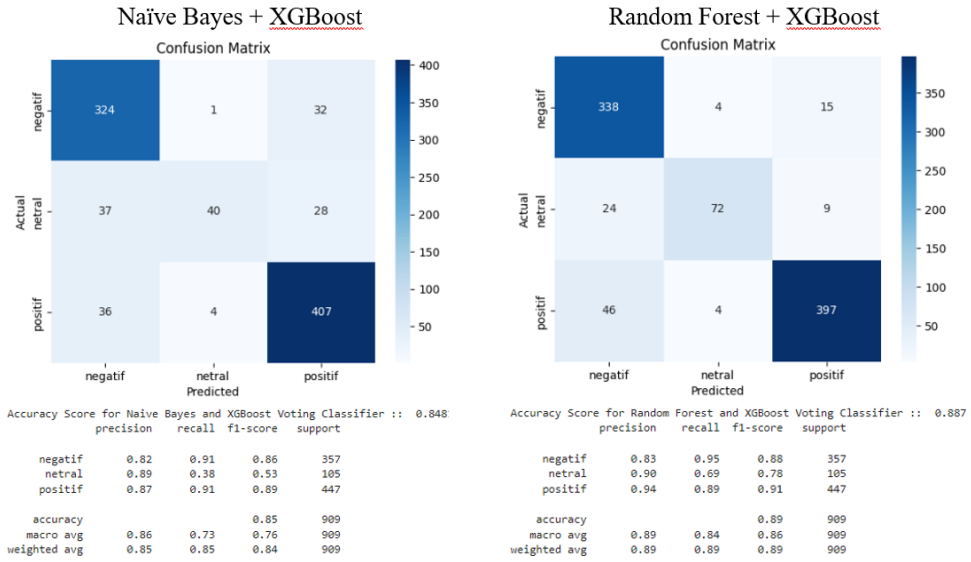

Figure 4. Comparison of Confusion Matrix 2 Algorithms + XGBoost

Based on the confusion matrix and performance metrics displayed, the testing of the Naive Bayes model combined with XGBoost shows a significant improvement in performance for sentiment classification related to the relocation of Indonesia's new capital city (IKN). The model achieved an accuracy of 84.8% or 85%, which is an improvement over the previous Naive Bayes model. Out of 909 samples, the model correctly classified 324 negative sentiments, 40 neutral sentiments, and 407 positive sentiments. The highest precision was found in the neutral class (0.89), though the recall for this class remains low (0.38), indicating that while the model is accurate in predicting the neutral class, many samples in this class are still not detected. The highest F1-score was achieved in the positive class (0.89), indicating a good balance between precision and recall. The increase in recall for the negative and positive classes shows that the model can correctly identify more samples from these classes after applying XGBoost. Overall, the combination of Naive Bayes and XGBoost strengthens the model by providing more accurate and balanced predictions, especially in handling data imbalance and improving detection in minority classes.

Based on the confusion matrix and performance metrics for the combined Random Forest and XGBoost model, it is evident that this model delivers superior performance with an accuracy of 88.7% in sentiment classification related to the relocation of Indonesia's new capital city (IKN). Out of 909 samples, the model correctly classified 338 negative sentiments, 72 neutral sentiments, and 397 positive sentiments. The highest precision was found in the neutral class (0.90), indicating that the model is highly accurate in predicting this class, while the highest recall was achieved in the negative class (0.95), signifying that the model can correctly detect most samples from this class. The highest F1-score was found in the positive class (0.91), showing that the model can provide a good balance between precision and recall for predictions in this class. The combination of XGBoost with Random Forest enhances predictive power by providing more balanced and accurate results, particularly in dealing with data imbalance. These results indicate that the use of this ensemble technique is very effective in enhancing model performance, making predictions more reliable and accurate on complex data.

Algorithms and SMOTE

This test aims to evaluate the performance of the Naive Bayes and Random Forest algorithms after applying the Synthetic Minority Over-Sampling Technique (SMOTE) in the classification of public sentiment related to the relocation of Indonesia's new capital city (IKN). SMOTE is used to address data imbalance by adding synthetic samples to the minority class, which is expected to increase the model's accuracy and ability to detect underrepresented sentiments. This analysis will show how the application of SMOTE affects the performance of the Naive Bayes and Random Forest models, illustrated through a confusion matrix to provide a detailed overview of the improvement in sentiment class predictions.

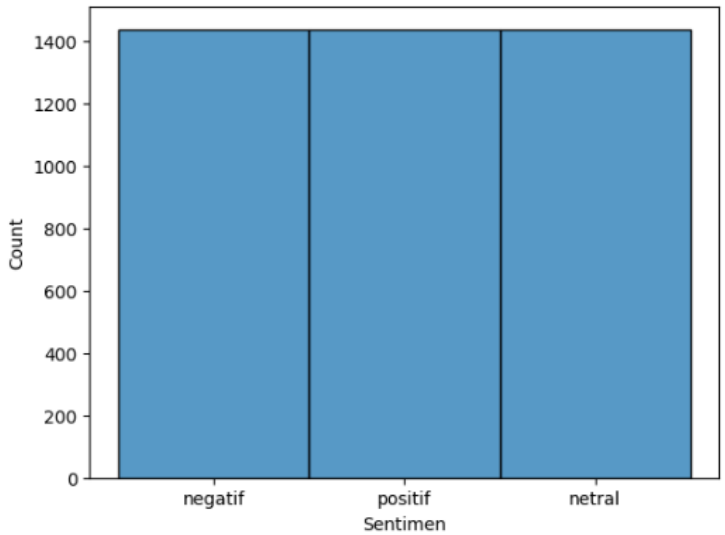

Figure 5. Data Balancing Results With SMOTE

The graph above shows the results of applying the Synthetic Minority Over-Sampling Technique (SMOTE) to sentiment data related to the relocation of the IKN in Indonesia. Before applying SMOTE, the data showed an imbalance with

varying numbers of samples for each sentiment category. However, after applying SMOTE, the data became balanced with the same number for each sentiment class: negative, positive, and neutral, each amounting to around 1,400 samples. This shows that SMOTE has succeeded in adding synthetic samples to the minority class, so that the data distribution becomes more balanced. With a balanced distribution, machine learning models are expected to improve performance in detecting previously underrepresented sentiments and reducing bias towards the majority class, thereby increasing the accuracy and ability of the model to make predictions.

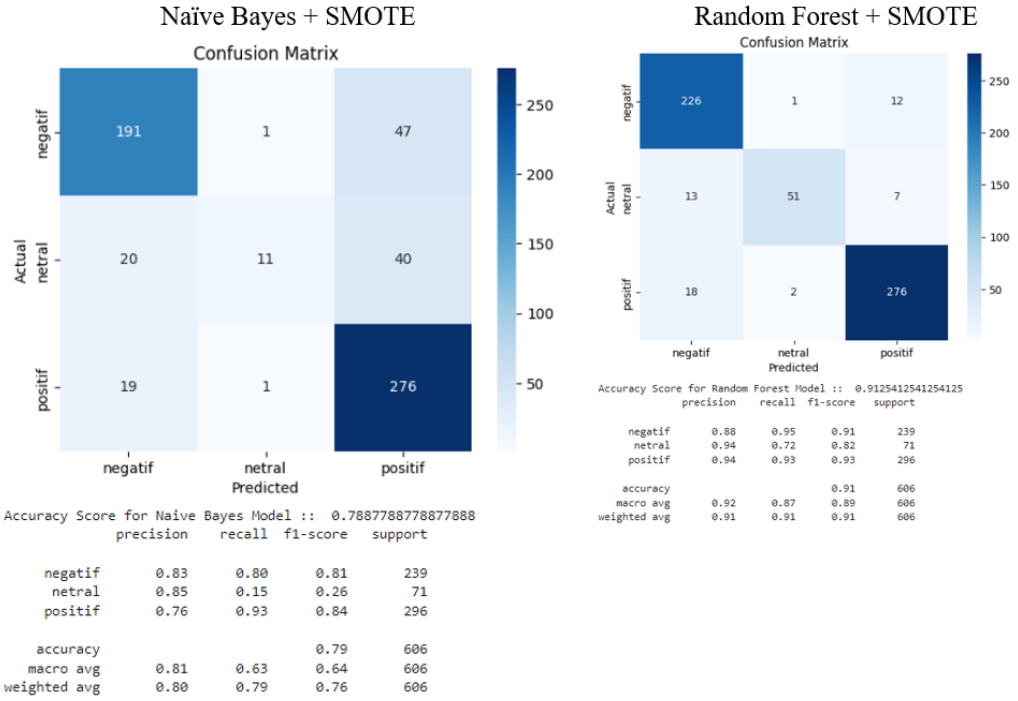

Figure 6. Comparison of Confusion Matrix 2 Algorithms + SMOTE

Based on the confusion matrix and performance metrics for the Naive Bayes model that has been applied with SMOTE, it can be seen that this model has an accuracy of 78.9%. The application of SMOTE aims to balance the number of samples in each sentiment class, which is expected to improve the model's ability to detect minority classes. Of the 606 samples, the model successfully classified 191 negative sentiments, 11 neutral sentiments, and 276 positive sentiments correctly. The highest precision was obtained in the neutral class (0.85), but the recall for this class was low (0.15), indicating that although the model is selective in predicting the neutral class, many samples from this class are not detected. The highest F1-score was achieved in the positive class (0.84), indicating a good balance between precision and recall. Although the application of SMOTE successfully improved the balance of the data, the results show that the model's performance can still be further improved, especially in handling neutral class predictions. This indicates that although SMOTE helps, there is a need for additional methods to improve the detection of underrepresented classes. Based on the confusion matrix and performance metrics for the Random Forest model that has been applied with SMOTE, it can be seen that this model achieves excellent

accuracy of 91.3% in classifying sentiments related to the relocation of the Indonesian Capital City (IKN). From 606 samples, the model successfully classified 226 negative sentiments, 51 neutral sentiments, and 276 positive sentiments correctly. The highest precision was achieved in the neutral class (0.94), indicating that the model is very accurate in predicting samples from this class, while the highest recall was in the negative class (0.95), indicating that the model was able to detect most of the negative samples correctly. The highest F1-score was 0.93, for both the positive and negative classes, indicating an excellent balance between precision and recall in both classes. The application of SMOTE successfully balanced the data, which contributed to improving the detection of minority classes and improving the overall performance of the model. These results indicate that the combination of Random Forest and SMOTE is very effective in handling data imbalance, making the model more reliable and accurate in its predictions.

Algorithm, SMOTE and XGBoost

This test aims to evaluate the performance of the combination of the Naive Bayes algorithm with the SMOTE and XGBoost techniques in classifying public sentiment related to the relocation of the Indonesian Capital City (IKN) in Indonesia. By using SMOTE to balance the data and XGBoost to improve model accuracy through boosting, this test is expected to produce a significant increase in sentiment detection, especially for minority classes. The analysis will be carried out using a confusion matrix, which provides a detailed picture of the improvement in model performance from the application of these two techniques, as well as assessing how effective the combination of these methods is in improving overall classification results.

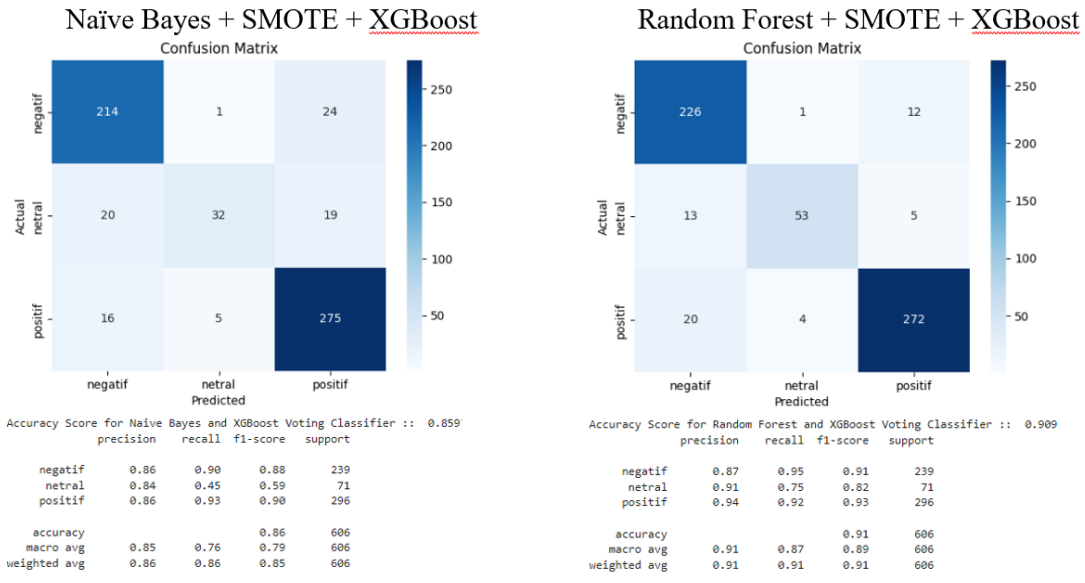

Figure 7. Comparison of Confusion Matrix 2 Algorithms + SMOTE + XGBoost

Based on the confusion matrix and performance metrics displayed, the combination of the Naive Bayes algorithm with SMOTE and XGBoost shows a significant performance improvement in the classification of sentiment related to

the relocation of the Indonesian Capital City (IKN). This model achieved an accuracy of 85.9%, indicating a clear improvement compared to previous testing. Out of 606 samples, the model successfully classified 214 negative sentiments, 32 neutral sentiments, and 275 positive sentiments correctly. The highest precision and recall were achieved in the positive class (0.86 and 0.93), indicating the model's ability to detect and predict this class very well. Although the neutral class still shows challenges with a recall of 0.45, the f1-score for this class increased to 0.59, indicating an improvement from previous results. The combination of SMOTE and XGBoost successfully corrected bias towards the majority class and improved detection of the minority class, making the model more accurate and balanced in its predictions. Overall, the use of XGBoost in this combination successfully improved the classification capabilities of Naive Bayes, providing more accurate and reliable results.

Based on the confusion matrix and performance metrics for the combination model of Random Forest, SMOTE, and XGBoost, this model shows a very high accuracy of 91.9% in classifying sentiments related to the relocation of the IKN. From a total of 606 samples, this model successfully classified 226 negative sentiments, 53 neutral sentiments, and 272 positive sentiments correctly. The highest precision was achieved in the positive class (0.94), indicating very good accuracy in predicting this class, while the highest recall was found in the negative class (0.95), indicating that the model is very effective in detecting most negative samples. The highest F1-score was 0.93, indicating an optimal balance between precision and recall, especially in the positive and negative classes. The application of SMOTE helps balance the data, while the use of XGBoost improves the predictive power of the model by increasing the accuracy and stability of the prediction results. Overall, this combination of techniques has proven to be very effective in improving the model's ability to handle imbalanced data and provide more reliable and accurate predictions.

Comparison Results

The results of this comparison will describe the performance comparison between various combinations of algorithms and techniques that have been applied in sentiment classification related to the relocation of the IKN. By comparing performance metrics such as accuracy, precision, recall, and f1-score, this analysis aims to identify which approach provides the best results in handling complex and imbalanced data. The results of this comparison will provide important insights into the effectiveness of each model, help in choosing the most efficient and reliable solution for sentiment analysis on social media platforms, and provide direction for future model development.

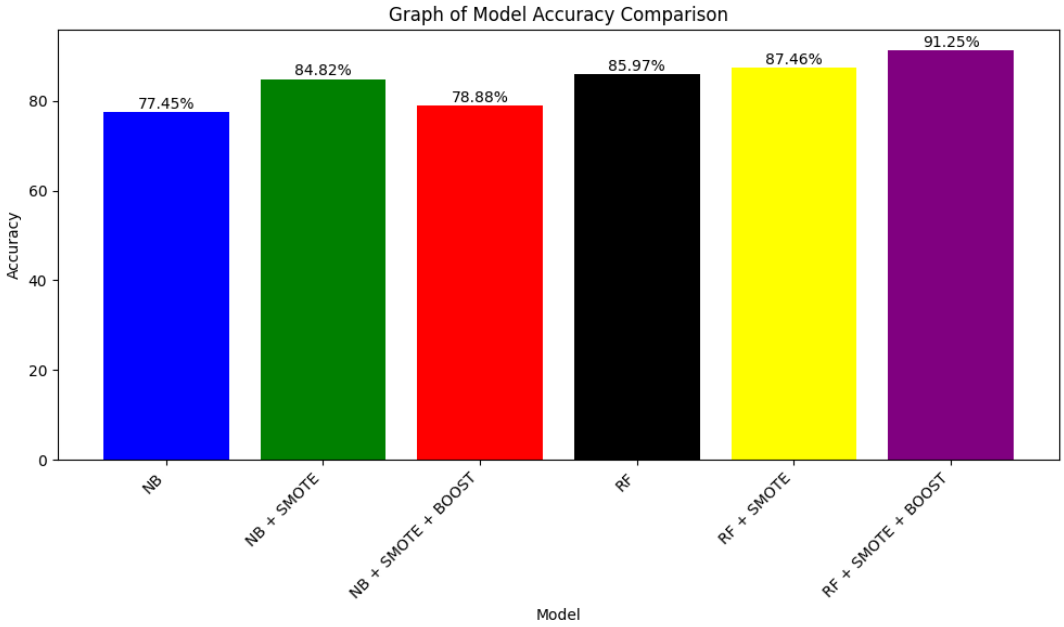

Figure 8. Model Comparison Chart

The comparison graph of accuracy values shows the performance of various combinations of models and techniques applied in the classification of sentiment related to the relocation of the IKN. The Naive Bayes (NB) model has a base accuracy of 77.45%, which increases to 84.82% when combined with XGBoost, indicating that boosting can significantly improve accuracy. The application of SMOTE to Naive Bayes slightly increases the accuracy to 78.88%, and further increases to 85.97% when SMOTE is combined with XGBoost, indicating the benefits of SMOTE in balancing data. The Random Forest (RF) model shows a base accuracy of 87.46%, higher than Naive Bayes. The combination of RF with XGBoost increases the accuracy to 88.78%, highlighting the effectiveness of the boosting technique. The best combination is achieved with Random Forest applied with SMOTE, achieving the highest accuracy of 91.25%, indicating that data balancing with SMOTE is very effective. Although RF+SMOTE+XGBOOST is slightly lower at 90.92%, it still indicates that the combination of these techniques provides excellent performance. Overall, the use of SMOTE and XGBoost in combination with Random Forest provides the most optimal results, strengthening the model to provide more accurate and reliable predictions.

D. Conclusion

This study has explored various combinations of algorithms and techniques to improve the accuracy of sentiment classification related to the relocation of the IKN in Indonesia. The combination of algorithms used includes Naive Bayes, Random Forest, SMOTE, and XGBoost. The results showed that the combination of Random Forest with SMOTE gave the best results with an accuracy of 91.25%. The application of SMOTE successfully balanced the unbalanced dataset, improving the model's ability to detect sentiment from the minority class. In the comparative results, the Naive Bayes (NB) model had an accuracy of 77.45%, which increased to

84.82% with XGBoost, and 78.88% with SMOTE. The combination of NB with SMOTE and XGBoost reached 85.97%. The Random Forest (RF) model itself reached 87.46% and increased to 88.78% with XGBoost. Interestingly, Random Forest with SMOTE showed the highest accuracy of 91.25% compared to the combination of RF, SMOTE, and XGBoost which had an accuracy of 90.92%, because the addition of XGBoost can add complexity that slightly reduces the generalization ability of the model. Further research opportunities can be focused on developing more adaptive models by incorporating other techniques such as deep learning or using transformer models to handle more complex text data. Researchers can further tune the parameters for Random Forest and XGBoost to optimize model performance. In addition, future studies can expand the scope of the data by combining data from various social media platforms or using real-time data to capture the dynamics of ever-changing sentiment. With a more holistic approach, future research can provide deeper insights into public perception and help in formulating more effective policies, as well as identifying other factors that can affect the effectiveness of predictive models.

E. Acknowledgment

Acknowledgments are addressed to Universitas Hang Tuah Pekanbaru for its support and contribution through research funding and community service in 2024.

F. References

[1] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, 2021, doi: 10.29207/resti.v5i3.3067.

[2] N. Widjiyati, "Implementasi Algoritme Random Forest Pada Klasifikasi Dataset Credit Approval," *J. Janitra Inform. dan Sist. Inf.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.25008/janitra.v1i1.118.

[3] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Sci. Financ. Econ.*, vol. 3, no. 4, pp. 354–379, 2023.

[4] E. Priyanto, E. I. Sela, L. A. Latumakulita, and N. Islam, "Decision Tree C4.5 Performance Improvement using Synthetic Minority Oversampling Technique (SMOTE) and K-Nearest Neighbor for Debtor Eligibility Evaluation," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 373–381, 2023, [Online]. Available: https://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/1676

[5] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[6] X. Dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," *2021 IEEE Int. Conf. Consum. Electron. Comput. Eng. ICCECE 2021*, no. Iccece, pp. 480–483, 2021, doi: 10.1109/ICCECE51280.2021.9342304.

[7] R. K. Septiani, S. Anggraeni, and S. D. Saraswati, "Klasifikasi Sentimen Terhadap Ibu Kota Nusantara (IKN) pada Media Sosial Menggunakan Naive Bayes," *Teknika*, vol. 16, no. 2, pp. 245–254, 2022.

[8] M. K. Anam, T. A. Fitri, A. Agustin, L. Lusiana, M. B. Firdaus, and A. T. Nurhuda, "Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 290–302, 2023, [Online]. Available: https://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/1590

[9] M. Yasir and R. Suraji, "Perbandingan Metode Klasifikasi Naive Bayes, Decision, Tree, Random Forest Terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube," *J. Cahaya Mandalika*, vol. 3, no. 2, pp. 180–192, 2023.

[10] A. Taufiq Akbar, H. Prapcoyo, and R. Husaini, "SMOTE and K-Means Preprocessing for Classification by Logistic Regression on Pima Indian Diabetes Dataset Prapemrosesan Menggunakan SMOTE dan K-means untuk Klasifikasi Regresi Logistik pada Data Pima Indian Diabetes," *J. Inform. dan Teknol. Inf.*, vol. 20, no. 2, pp. 238–249, 2023, doi: 10.31515/telematika.v20i2.9676.

[11] I. P. A. P. Widiarta, R. Dwiyansaputra, and A. Aranta, "Analisis Sentimen Masyarakat Terhadap Kebijakan Penerapan Ppkm Di Media Sosial Twitter Dengan Menggunakan Metode Xgboost," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 5, no. 2, pp. 154–163, 2023, doi: 10.29303/jtika.v5i2.342.

[12] H. Christanto *et al.*, "Analisis Perbandingan Decision Tree, Support Vector Machine, dan Xgboost dalam Mengklasifikasi Review Hotel Trip Advisor," *J. Teknol. Inform. dan Komput.*, vol. 9, no. 1, pp. 306–319, 2023, doi: 10.37012/jtik.v9i1.1429.

[13] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. Dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: 10.55606/jurrimipa.v2i2.1470.

[14] D. F. Wicaksono, R. S. Basuki, and D. Setiawan, "Peningkatan Performa Model Machine Learning XGBoost Classifier melalui Teknik Oversampling dalam Prediksi Penyakit AIDS," vol. 8, no. April, pp. 736–747, 2024, doi: 10.30865/mib.v8i2.7501.

[15] M. Novo-Lourés, R. Pavón, R. Laza, D. Ruano-Ordas, and J. R. Méndez, "Using natural language preprocessing architecture (NLPA) for big data text sources," *Sci. Program.*, vol. 2020, 2020, doi: 10.1155/2020/2390941.

[16] W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36–44, 2021, doi: 10.25008/ijadis.v2i1.1216.

[17] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 776–784, 2022, doi: 10.11591/ijece.v12i1.pp776-784.

[18] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022, doi:

10.1177/1094428120971683.
[19] R. Kalaivani and R. Marivendan, "The effect of stop word removal and stemming in datapreprocessing," *Ann. R.S.C.B*, vol. 25, no. 6, pp. 739–746, 2021.
[20] A. N. Ulfah, M. K. Anam, N. Y. Sidratul Munti, S. Yaakub, and M. B. Firdaus, "Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19," *JUITA J. Inform.*, vol. 10, no. 2, p. 209, 2022, doi: 10.30595/juita.v10i2.12308.
[21] Junadhi, Agustin, M. Rifqi, and M. K. Anam, "Sentiment Analysis of Online Lectures using K-Nearest Neighbors based on Feature Selection," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 3, pp. 216–225, 2022, doi: 10.23887/janapati.v11i3.51531.
[22] T. Colibazzi *et al.*, "Identifying Splitting Through Sentiment Analysis," *J. Pers. Disord.*, vol. 37, no. 1, pp. 36–48, 2023, doi: 10.1521/pedi.2023.37.1.36.
[23] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," *Procedia Comput. Sci.*, vol. 234, pp. 349–356, 2024, doi: 10.1016/j.procs.2024.03.014.
[24] M. Parzinger, L. Hanfstaengl, F. Sigg, U. Spindler, U. Wellisch, and M. Wirnsberger, "Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection — Using machine learning methods Random Forest and LASSO," *Build. Environ.*, vol. 223, no. February, p. 109313, 2022, doi: 10.1016/j.buildenv.2022.109313.
[25] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.
[26] J. Xu, Y. Jiang, and C. Yang, "Landslide Displacement Prediction during the Sliding Process Using XGBoost, SVR and RNNs," *Appl. Sci.*, vol. 12, no. 12, 2022, doi: 10.3390/app12126056.