ISSN [e]: 2579-4264 | DOI: https://doi.org/10.26760/jrh.V10i1.67-80 Maret 2026

Machine Learning Optimization on Social Media Sentiment Data for Data Balance Using N-GRAM

Rizka Milandga Milenio[1*], Jasman Pardede[1], Dea Kurniasih[1]

[1]Institut Teknologi Nasional, Bandung, Indonesia

Email: rizkamilandga@itenas.ac.id, jasman@gmail.com, dea.kurniasih@mhs.itenas.ac.id

25 Januari 2026 | Revised 2 Februari 2026 | Accepted 10 Februari 2026

ABSTRAK

Ketidakseimbangan kelas merupakan tantangan dalam klasifikasi sentimen pada data media sosial, yang menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas dan berkinerja buruk pada kelas minoritas. Penelitian ini mengusulkan pendekatan penyeimbangan data berbasis N-Gram untuk mengatasi masalah tersebut dan meningkatkan performa klasifikasi. Tiga model machine learning, yaitu XGBoost, Random Forest, dan Support Vector Machine (SVM), dievaluasi pada dataset yang tidak seimbang maupun seimbang menggunakan akurasi, presisi, recall, dan F1-score sebagai metrik evaluasi. Hasil eksperimen menunjukkan bahwa penyeimbangan data meningkatkan performa semua model tanpa menurunkan kemampuan generalisasi. SVM mencapai performa terbaik pada dataset seimbang dengan akurasi 0,86, presisi 0,87, recall 0,86, dan F1-score 0,86. XGBoost dan Random Forest juga menunjukkan peningkatan performa yang signifikan setelah penyeimbangan, menunjukkan kemampuan yang lebih baik dalam mendeteksi kelas minoritas. Secara keseluruhan, temuan ini menegaskan bahwa pendekatan penyeimbangan data berbasis N-Gram yang diusulkan efektif dalam mengurangi ketidakseimbangan kelas dan meningkatkan ketahanan serta keandalan model klasifikasi sentimen.

***Kata kunci**: klasifikasi sentimen, ketidakseimbangan kelas, n-gram, media sosial*

ABSTRACT

Class imbalance is a challenge in sentiment classification of social media data, often causing classification models to be biased toward majority classes and perform poorly on minority classes. This study proposes an N-Gram-based data balancing approach to address this issue and improve classification performance. Three machine learning models, namely XGBoost, Random Forest, and Support Vector Machine (SVM), were evaluated on both imbalanced and balanced datasets using accuracy, precision, recall, and F1-score as evaluation metrics. The experimental results demonstrate that data balancing consistently enhances performance across all models without degrading generalization capability. Among the evaluated methods, SVM achieves the best performance on the balanced dataset, reaching an accuracy of 0.86, precision of 0.87, recall of 0.86, and F1-score of 0.86. XGBoost and Random Forest also show substantial performance improvements after balancing, indicating improved detection of minority sentiment classes. Overall, the findings confirm that the proposed N-Gram-based data balancing approach effectively mitigates class imbalance and improves the robustness and reliability of sentiment classification models.

***Keywords**: Sentiment Classification, Class Imbalance, N-Gram, Social Media*

1. INTRODUCTION

The condition of imbalanced data causes the distribution of sample data for training to be unbalanced, so the classification will tend to ignore the class/label with fewer instances and focus on the class/label with a large number of instances. This will affect the performance of the classification model. A high accuracy value cannot be used as an indication of the model's performance because the model mostly only correctly classifies the majority class [1].

Data augmentation in sentiment analysis is used to enhance the robustness and generalization capability of sentiment prediction models. In this study, augmentation is also performed due to the imbalance in the amount of data between positive and negative label categories. Dataset with imbalanced label data will affect the construction of text classification models using machine learning. The impact of the imbalance in the number of classes (categories) in the dataset is the occurrence of misclassification of the minority class (with fewer data), which can affect the overall classification performance [2]. The data augmentation technique used in this study. The n-gram concept is adopted to enhance the diversity of the training data by predicting words that might appear in a specific context. N-grams, as a sequence of n elements, allow for modelling the relationships between words in the text. By using unigrams, bigrams, and trigrams [3].

Based on [4] The random forest algorithm is the best algorithm for this dataset from the test process review with an accuracy of 86% (85.76%), SVM and Xgboost each have an accuracy value of 86% (85.58%) and 84%. Then the second, Stemming can increase accuracy and reduce accuracy depending on the amount of data used. And finally, if you don't use stemming, the accuracy results will decrease with each result of 85% and the results of these three algorithms are slightly different, namely SVM 84.87%, XGBoost 84.60% and Random Forest 84.71%

The results were that classification using random forest obtained a testing accuracy of 79.22%, with using support vector classification gets a testing accuracy of 76.62%, using XGBoost gets a testing accuracy of 79.22% [5]. Furthermore, another study showed that the research findings indicate that the model accuracy is 75.17% (SVM), 84.06% (RF), and 83.17% (XGBoost). This result proves XGBoost is the most stable and optimal algorithm for sentiment classification tasks in the Access by KAI application [6]. To enable classification using machine learning algorithms, the extracted N-gram features are converted into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme, which serves as input for the classification models.

In addition to addressing class imbalance, sentiment labeling quality also becomes a critical issue in Indonesian sentiment analysis. This study initially employed the Indonesia Sentiment Lexicon (INSET) as an automated lexicon-based labeling approach due to its efficiency in assigning sentiment polarity labels to large-scale textual datasets. INSET contains predefined sentiment scores for Indonesian vocabulary and is commonly used for automatic sentiment classification. However, lexicon-based approaches often struggle to capture contextual meaning, sarcasm, implicit criticism, and domain-specific expressions commonly found in social media texts. Therefore, this study further evaluates the reliability of INSET by comparing its labeling results with manual human validation before performing classification experiments.

This study addresses that gap by comparing the performance of five text embedding models TF-IDF, Word2Vec, FastText, BERT, and GPT, Experimental results show that the combination of Gaussian Naive Bayes with GPT embeddings achieved the best performance in ambiguous sentence classification, with a recall of 71% and an F1- score of 60%. Meanwhile, the combination of TF-IDF with bagging yielded the highest accuracy of 83% for unambiguous sentence classification. These findings highlight the critical role of selecting appropriate embedding and classification models to enhance accuracy in semantically ambiguous sentence classification for the Indonesian language [7].

Based on the discussion that has been outlined, data imbalance in text classification can cause bias in the classification model, where the majority class is more often correctly identified than the minority class. Therefore, data augmentation using the N-Gram method becomes one of the effective solutions in increasing the diversity of training data and addressing this imbalance issue. In addition to data augmentation, selecting appropriate classification algorithms plays a crucial role in improving model performance.

Based on previous studies, machine learning algorithms such as Support Vector Machine, Random Forest, and XGBoost have demonstrated strong performance in sentiment classification tasks. Based on the identified challenges of class imbalance and labeling reliability in Indonesian social media sentiment analysis, this study aims to: (1) evaluate the reliability of the Indonesia Sentiment Lexicon (INSET) compared with manual human labeling, (2) propose an N-Gram-based synthetic data balancing method to overcome class imbalance, and (3) compare the performance of XGBoost, Random Forest, and Support Vector Machine (SVM) models using TF-IDF feature representation on both imbalanced and balanced datasets. The proposed framework combines automated labeling evaluation, synthetic text augmentation using unigram, bigram, and trigram generation, and machine learning classification to improve sentiment classification robustness in Indonesian social media datasets.

2. METHODS

2.1 Dataset and Preprocessing

The dataset used in this study was collected through data crawling from the social media platform X (formerly Twitter). Based on figure 1, data retrieval was conducted using authenticated access and predefined keywords related to the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek). Only Indonesian-language tweets were retained to ensure linguistic consistency. The collection period spans from January 1, 2010, to January 1, 2025, resulting in a total of 3,434 textual instances. The raw data were stored in CSV format, with the primary textual content extracted from the full_text attribute.

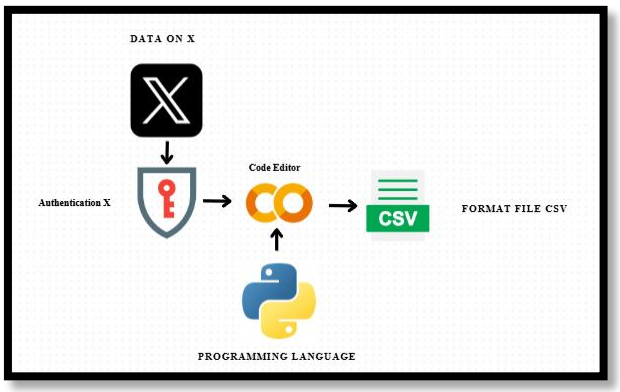


Figure 1. Crawling dataset process

Following data acquisition, based on figure 2, a structured preprocessing pipeline was applied to enhance data quality and suitability for machine learning classification. The preprocessing stages consisted of case folding, normalization, cleaning, tokenization, and stopword removal. Case folding converted all characters to lowercase to eliminate inconsistencies caused by letter casing. The normalization step addressed informal and non-standard word forms by mapping them to standardized Indonesian vocabulary references. Cleaning involved removing URLs, hashtags, mentions, punctuation marks, numeric characters, and other irrelevant symbols that do not contribute to semantic meaning.

Subsequently, tokenization was performed to segment each tweet into individual lexical units. Stopword removal was then applied to eliminate high-frequency functional words (e.g., conjunctions and common particles) that carry limited discriminative information for sentiment classification.

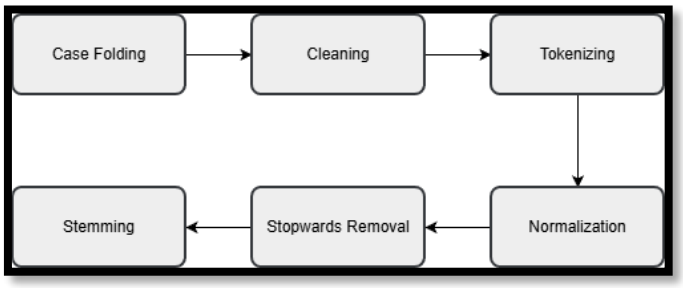


Figure 2. Pre-processing

After preprocessing, the dataset was manually labeled into three sentiment categories: positive, negative, and neutral. The class distribution before balancing consisted of 1,860 positive, 561 negative, and 1,013 neutral instances, indicating a substantial class imbalance that motivated the implementation of data balancing techniques in the subsequent stage of the study.

2.2 Data Balancing with N-Gram

In order to address the substantial class imbalance observed in the dataset, a synthetic data generation approach based on the N-Gram language modeling technique was implemented. Prior to balancing, the dataset distribution consisted of 1,860 positive instances, 561 negative instances, and 1,013 neutral instances. Such imbalance can bias classification models toward the majority class, thereby reducing sensitivity to minority sentiments. To mitigate this issue, synthetic text instances were generated for the minority classes until all sentiment categories reached equal representation.

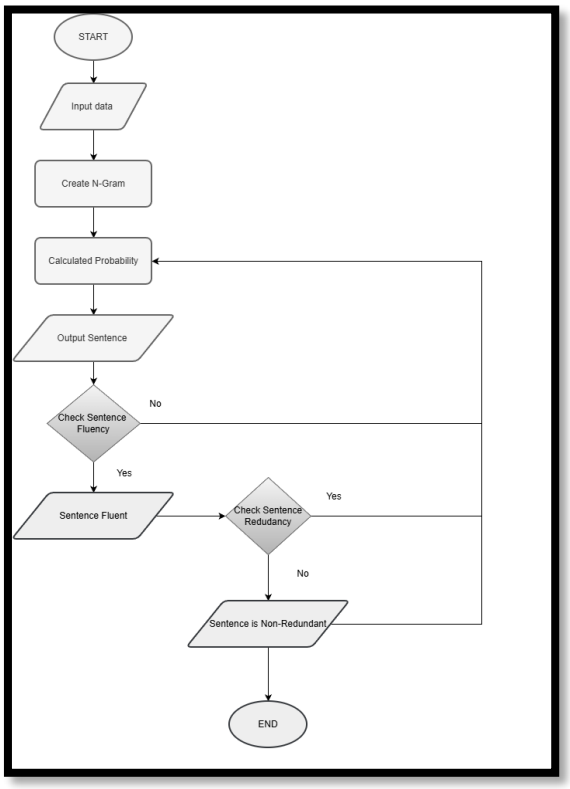


Figure 3. Flowchart N-Gram

Based on figure 3, the N-Gram approach models textual sequences by estimating the probability of a word given its preceding context. In this study, unigram, bigram, and trigram structures were utilized to capture local lexical dependencies within the corpus. Word transition probabilities were computed based on frequency statistics derived from the original training data. Synthetic sentences were then generated by sampling word sequences according to these learned probability distributions, allowing the preservation of contextual coherence while introducing lexical variation.

The augmentation process was applied exclusively to the negative and neutral classes, as the positive class already contained sufficient instances. Synthetic data were generated iteratively until each class contained 1,860 instances, resulting in a fully balanced dataset. Quality control was conducted to ensure that generated sentences did not duplicate existing original or synthetic texts. Additionally, sentence length distributions were monitored to maintain consistency with the natural characteristics of the corpus.

The resulting balanced dataset demonstrated improved representational parity across sentiment classes, thereby reducing prior probability bias in model training. This N-Gram-based augmentation strategy provides a lightweight and interpretable alternative to more complex oversampling methods, while effectively enhancing minority class representation for subsequent classification tasks.

2.3 Feature Representation and Classification Models

Strategy to transform textual data into machine-readable form, the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme was employed as the numerical feature representation method. TF-IDF quantifies the importance of a term within a document relative to its frequency across the entire corpus. This approach emphasizes discriminative terms while down-weighting common words, thereby

improving the separability of sentiment-related features. The preprocessed tweets were vectorized using TF-IDF to generate high-dimensional sparse feature matrices suitable for supervised learning algorithms.

Three classification models were implemented to evaluate sentiment prediction performance under imbalanced and balanced conditions: Extreme Gradient Boosting (XGBoost), Random Forest, and Support Vector Machine (SVM). These models were selected due to their robustness in handling high-dimensional textual data and their demonstrated effectiveness in sentiment analysis tasks.

XGBoost is an ensemble learning algorithm based on gradient boosting of decision trees, designed to optimize predictive performance through iterative error correction and regularization. Random Forest is a bagging-based ensemble method that constructs multiple decision trees and aggregates their outputs to reduce variance and improve generalization. SVM is a margin-based classifier that identifies an optimal hyperplane to maximize class separation in high-dimensional space, making it particularly suitable for sparse text representations.

Hyperparameter tuning was conducted using grid search to identify optimal configurations for each model. Model training and evaluation were performed on both the original imbalanced dataset and the N-Gram-balanced dataset to assess the impact of data balancing on classification performance. This comparative framework enables systematic analysis of how feature representation and model architecture interact with class distribution in multi-class sentiment classification.

2.4 Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC–AUC). Accuracy measures the overall proportion of correctly classified instances across all sentiment classes, as defined in Equation (1). However, given the multi-class and imbalanced nature of the dataset, Accuracy alone is insufficient to reflect classification quality.

Precision and Recall were therefore computed to assess class-level performance. Precision indicates the proportion of correctly predicted positive instances among all predicted positives, as shown in Equation (2), while Recall measures the proportion of correctly identified instances among actual positives, as defined in Equation (3). The F1-score, defined as the harmonic mean of Precision and Recall, is presented in Equation (4) provides a balanced measure of classification effectiveness, particularly in imbalanced settings.

To further evaluate discriminative capability, ROC curves and macro-averaged AUC values were calculated. The macro-averaging strategy ensures equal weight across classes, enabling comprehensive assessment of model performance under both imbalanced and balanced data conditions. Below are the detailed formulas for each evaluation metric used in this study.

1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
2) $Precision = \frac{TP}{TP+FP}$
3) $Recall = \frac{TP}{TP+FN}$
4) $F1 - Score = \frac{Precision\ x\ Recall}{Precision+Recall}$

3. RESULT AND DISCUSSION

3.1 Labelling Validation and Reliability

To ensure the reliability and validity of the sentiment labels used in this study, a comparative evaluation was conducted between automated lexicon-based labeling using the Indonesia Sentiment Lexicon (INSET) and manual annotation performed by human validators. This assessment aimed to identify labeling inconsistencies and examine the limitations of rule-based sentiment classification in capturing contextual meaning. A representative subset of the dataset was analyzed to measure agreement levels and error patterns. The comparison provides empirical evidence regarding annotation quality and establishes a reliable foundation for subsequent model training and performance evaluation.

Table 1. Different labelling inset and validators

Mistake by Indonesia Sentiment Lexicon (INSET)				
Teks Original	**Teks Synthetic**	**Lexicon**	**Validator**	**Corrected Label**
kemendikbudristek dipecah dosen unnes: ada potensi hambatan komunikasi dan administrasi https://t.co/a3tsfqzaf6 #temponasional	*kemendikbudristek pecah dosen unnes potensi hambat komunikasi administrasi*	Neutral	Negative	Negative
Mistake By Human Validators				
Teks Original	**Teks Synthetic**	**Lexicon**	**Validators (Human Erorr)**	**Corrected Label**
selain itu kemendikbudristek juga mendapatkan kritik atas program-program unggulan mereka seperti mbkm iisma dan permasalahan lainnya. https://t.co/ov5hjj4ovz	*kemendikbudristek kritik program program unggul mbkm iisma masalah*	Negative	Neutral	Negative

An evaluation was conducted by comparing sentiment labeling results between the Indonesia Sentiment Lexicon (INSET) and human validators. Based on table 1, this comparison aimed to identify differences in sentiment labels and to analyze the sources of labeling errors, whether originating from the lexicon-based system or from human validators. The evaluation results indicate that in several cases, INSET produced sentiment labels that differed from those assigned by human validators. These discrepancies generally occurred in texts containing implicit criticism or negative context that was not explicitly represented by sentiment-bearing words in the lexicon.

Table 2. Result comparing labelling

Description	**Count**	**Precentage**
Success rate INSET	382/687	55,89%
Success rate Validator	685/687	99,70%

The relatively low performance of INSET (55.89%) indicates that lexicon-based sentiment labeling has significant limitations when applied to Indonesian social media data. First, INSET relies heavily on individual word polarity and cannot effectively interpret contextual sentiment shifts caused by sarcasm, irony, negation, or implicit criticism. Second, social media texts frequently contain informal expressions, abbreviations, slang, and topic-specific vocabulary related to educational policy issues that may not exist in the INSET dictionary. Third, sentiment polarity in public policy discussions is often expressed indirectly, causing lexicon-based systems to misclassify neutral-looking sentences that

actually contain negative sentiment. These findings justify the use of manual validation as a more reliable labeling strategy for training machine learning models.

To quantitatively assess labeling reliability, as a result of table 2, a stratified evaluation was conducted on 20% of the total dataset, corresponding to 687 instances out of 3,434 collected tweets. The sentiment labels assigned by the Indonesia Sentiment Lexicon (INSET) were compared against manual annotations provided by human validators. The evaluation results indicate that INSET correctly classified 382 out of 687 instances, yielding an accuracy rate of 55.89%. In contrast, human validators achieved 685 correct labels, corresponding to a 99.70% agreement level, with only two identified labeling errors. These findings demonstrate a substantial discrepancy between automated lexicon-based labeling and manual annotation, suggesting that INSET struggles to capture contextual nuances, implicit criticism, and sentiment polarity shifts within Indonesian-language social media texts.

3.2 Effect of N-Gram Data Balancing

Based on figure 4, the initial dataset exhibited a substantial class imbalance across the three sentiment categories. Prior to augmentation, the dataset consisted of 1,860 positive instances, 561 negative instances, and 1,013 neutral instances. This distribution indicates a dominant positive class, with the negative class representing less than one-third of the majority class. Such imbalance introduces prior probability bias during model training, increasing the likelihood that classifiers favor the majority class while underperforming on minority categories. As shown in the distribution visualization, the disparity between classes is pronounced, particularly for the negative sentiment category. This imbalance provides the primary motivation for implementing a data balancing strategy to improve representational parity and reduce classification bias.

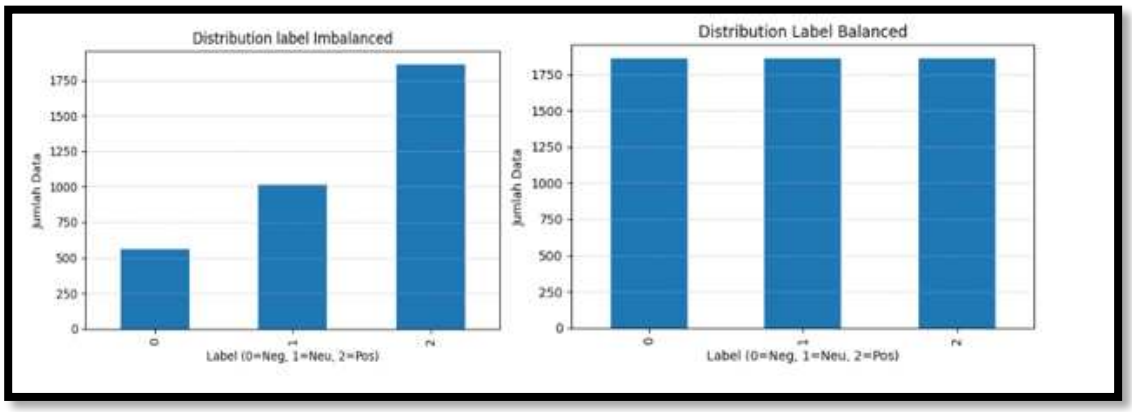


Figure 4. Distribution Label

Following the application of the N-Gram-based synthetic augmentation technique, the dataset achieved a fully balanced distribution. Synthetic instances were generated exclusively for the minority classes (negative and neutral) until each sentiment category reached 1,860 instances. As illustrated in the post-balancing distribution, all three classes now contain equal representation. This transformation eliminates skewed class priors and creates a more statistically uniform training environment, allowing classification models to learn sentiment boundaries without being disproportionately influenced by majority-class frequency.

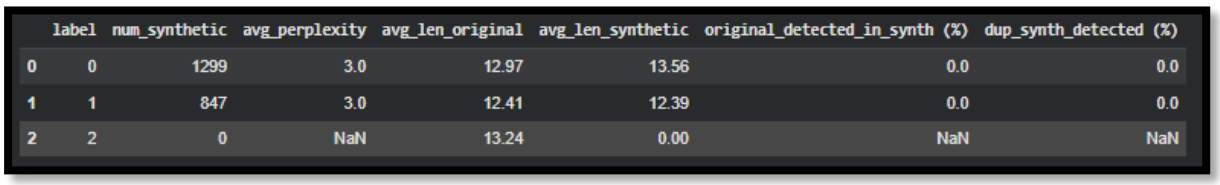
Table 3. Sentence Synthetic

Sentence Original	Sentence Synthetic	Label
kemendikbudristek hapus ekstrakurikuler pramuka kini pertimbangkan masuk kurikulum merdeka	*kemendikbudristek hapus ekstrakurikuler pramuka timbang masuk kurikulum merdeka tingkat*	*Neutral*
buku cabul masuk kurikulum muhammadiyah desak kemendikbudristek cabut buku panduannya	*buku cabul masuk kurikulum muhammadiyah desak kemendikbudristek ri sanksi pecat*	*Negative*

Based on table 3, beyond numerical balancing, the quality of synthetic data plays a critical role in determining the effectiveness of augmentation. The generated sentences were evaluated to ensure structural plausibility and semantic consistency with the original corpus. No duplication was detected between synthetic and original instances, nor among synthetic samples themselves, confirming that the augmentation process introduced genuine lexical variation rather than replication. Additionally, sentence length distributions remained consistent with the characteristics of the natural dataset, indicating that the N-Gram generation process preserved realistic textual patterns.

The evaluation further indicates that synthetic data were successfully generated for the negative and neutral classes, while no augmentation was required for the positive class due to its already sufficient representation. The inability to generate additional positive samples suggests that the N-Gram model primarily functions as a compensatory mechanism for underrepresented categories rather than an indiscriminate text generator.

Overall, the N-Gram balancing strategy effectively transformed an imbalanced sentiment dataset into a fully balanced corpus while maintaining textual diversity and structural integrity. This adjustment directly addresses class distribution bias and establishes a controlled experimental setting for evaluating classification performance under balanced conditions. The subsequent improvement in model metrics, discussed in the following section, provides empirical confirmation that the balancing process meaningfully enhances minority class representation and overall classification robustness.



	label	num_synthetic	avg_perplexity	avg_len_original	avg_len_synthetic	original_detected_in_synth (%)	dup_synth_detected (%)
0	0	1299	3.0	12.97	13.56	0.0	0.0
1	1	847	3.0	12.41	12.39	0.0	0.0
2	2	0	NaN	13.24	0.00	NaN	NaN

Figure 5. Evaluation N-Gram

Based on figure 5, the synthetic data augmentation process was successfully performed on the Negative and Neutral classes, with a considerable amount of synthetic data and good sentence quality. This is indicated by a low perplexity value, synthetic sentence lengths comparable to the original data, and no duplication found between the synthetic with original data and also synthetic with synthetic.
However, in the Positive class, the augmentation process did not generate synthetic data. This indicates a limitation of the model in forming new sentence variations for that class, which is likely due to the characteristics of the data or the distribution of language patterns in the Positive class. Overall, the generated synthetic data is of good quality and suitable for supporting the data balancing process in the classification stage.

3.3 Comparative Model Performance (Imbalanced vs Balanced)

The comparative evaluation of XGBoost, Random Forest, and Support Vector Machine (SVM) reveals consistent performance improvements after applying N-Gram-based data balancing. On the imbalanced dataset, all models exhibit moderate classification capability, with noticeable limitations in minority class detection. XGBoost achieved an accuracy of 0.69 and an F1-score of 0.64, while Random Forest recorded an accuracy of 0.6366 and an F1-score of 0.5984. SVM performed slightly better among the three, reaching an accuracy of 0.70 and an F1-score of 0.66. As table 4 shown, these results reflect the influence of skewed class priors, where models tend to optimize majority-class predictions at the expense of minority categories.

Table 4. comparative performance of classification models with best configuration

Model	Data Condition	Accuracy	Precision	Recall	F1-Score	Training Loss	Validation Loss
XGBoost	Imbalanced	0.69	0.63	0.65	0.64	0.85	0.85
XGBoost	Balanced	0.76	0.77	0.76	0.76	0.88	0.64
Random Forest	Imbalanced	0.64	0.60	0.63	0.60	1.00	1.00
Random Forest	Balanced	0.71	0.72	0.71	0.71	1.00	1.00
SVM	Imbalanced	0.70	0.66	0.68	0.66	0.89	0.63
SVM	Balanced	0.86	0.87	0.86	0.86	0.99	0.84

After balancing, performance improvements are observed across all evaluation metrics. XGBoost's accuracy increased to 0.7599 with an F1-score of 0.761. Random Forest improved to an accuracy of 0.7115 and an F1-score of 0.7127. The most substantial gain is observed in SVM, which achieved an accuracy of 0.86 and an F1-score of 0.86. The magnitude of improvement suggests that balancing significantly enhances class separability and reduces bias toward the majority sentiment class. Notably, the F1-score improvements indicate better harmonization between precision and recall, particularly for previously underrepresented categories.

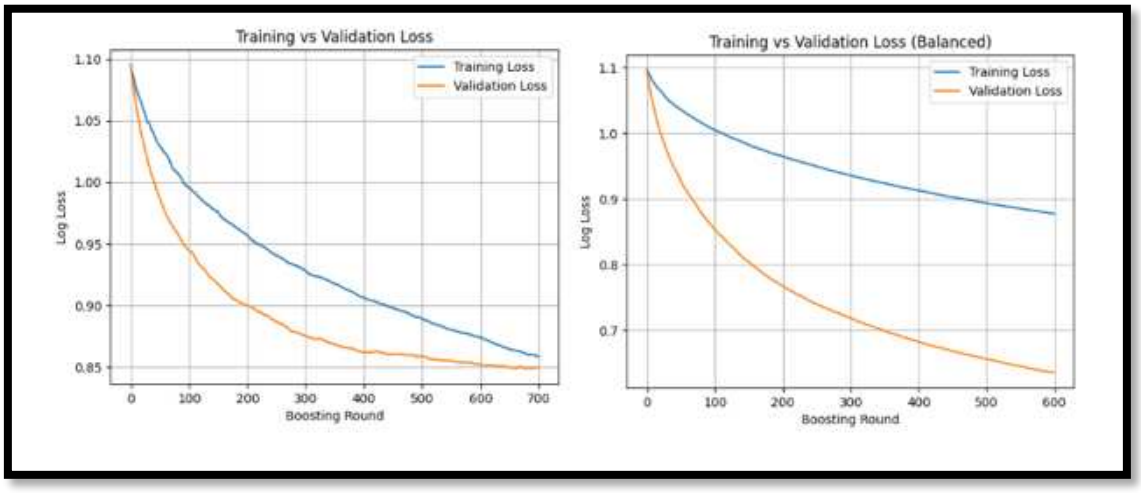


Figure 5. Visualization of training and validation loss (imbalanced vs balanced)

Based on figure 5. the training and validation loss curves further support the effectiveness of the balancing strategy. On the imbalanced dataset, loss values between training and validation remain relatively close for all models, indicating stable learning behavior without severe overfitting. However, predictive performance remains constrained by biased class representation.

On the balanced dataset, training and validation losses remain comparably aligned, demonstrating that the augmentation process does not introduce instability or excessive variance. For example, XGBoost shows reduced validation loss compared to the imbalanced condition, indicating improved generalization capability. Similarly, Random Forest maintains minimal loss divergence, while SVM exhibits a controlled increase in loss values alongside substantial gains in predictive performance. The small gap between training and validation losses confirms that performance improvements are not the result of overfitting but instead stem from enhanced class representation during learning.

Overall, comparative analysis demonstrates that N-Gram-based data balancing systematically improves classification robustness across different algorithmic paradigms, including boosting-based, bagging-based, and margin-based methods. Among the evaluated models, SVM benefits most significantly from balanced class distributions, suggesting that margin-based classifiers are particularly sensitive to prior probability distortion in multi-class sentiment tasks. These findings empirically validate the central hypothesis of this study: reducing class imbalance enhances minority class detection and strengthens overall predictive reliability without compromising generalization stability.

3.4 Evaluation and Discriminative Analysis

The evaluation results show that the Random Forest, SVM, and XGBoost models do not exhibit underfitting or overfitting on either imbalanced or balanced datasets, as indicated by the relatively small and stable gaps between training loss and validation loss. As of shown in figure 6. the random forest model, training and validation losses remain close on both the imbalanced dataset (1.017 and 1.036) and the balanced dataset (1.020 and 1.025), accompanied by improvements in accuracy from 0.6366 to 0.7115 and F1-score from 0.5984 to 0.7127 after data balancing.

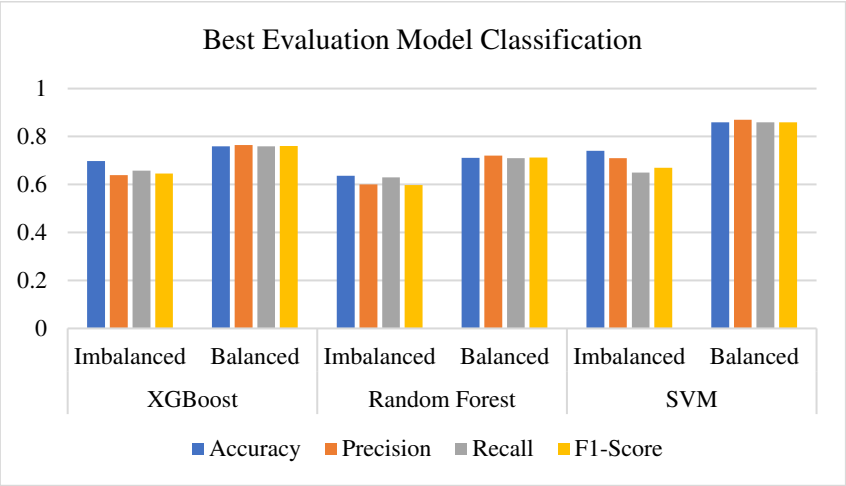


Figure 6. Best evaluation model classification

XGBoost demonstrates similar stable learning behavior, with training and validation losses of 0.858 and 0.849 on the imbalanced dataset and 0.877 and 0.635 on the balanced dataset, while accuracy increases from 0.69 to 0.7599 and F1-score from 0.64 to 0.761. The SVM model shows a controlled increase in training loss from 0.889 to 0.985 and validation loss from 0.633 to 0.838 after balancing, alongside a substantial performance improvement where accuracy and F1-score rise from 0.70 and 0.66 to 0.86 and 0.86, respectively. These results confirm that data balancing reduces bias caused by class imbalance

without introducing excessive variance, enabling all models to achieve better generalization and improved classification performance.

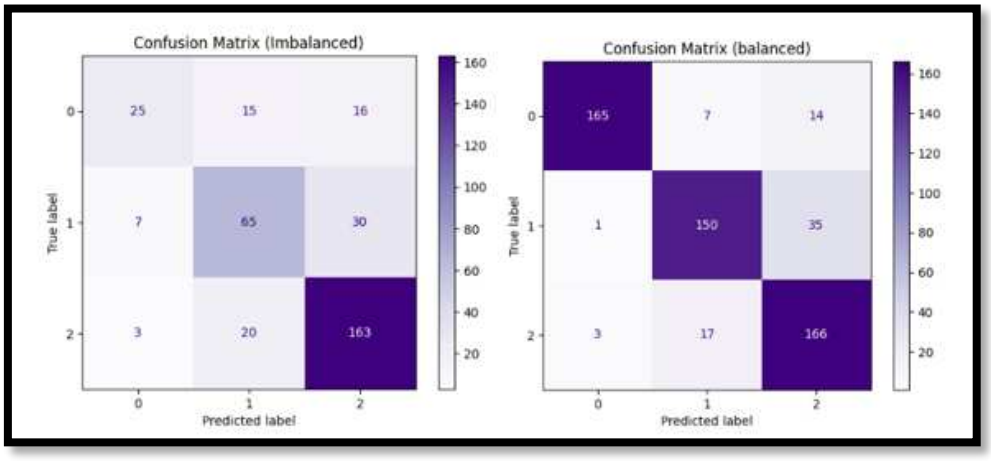


Figure 7. AUC ROC Imbalanced-Balanced

To further analyze the classification behavior at the class level, as of shown in figure 7, a confusion matrix is examined using the best-performing model as a representative example. This analysis aims to observe the distribution of correct and incorrect predictions for each sentiment class under imbalanced and balanced data conditions. Confusion matrices obtained from the imbalanced and balanced datasets. On the imbalanced dataset, the model shows a higher number of misclassifications, particularly for minority classes, as indicated by the relatively large off-diagonal values. For instance, class 0 is frequently misclassified as class 1 and class 2, while class 1 also exhibits notable confusion with class 2.

After applying data balancing, the confusion matrix demonstrates a significant improvement in classification performance across all classes. The number of correctly classified instances increases substantially, as reflected by the higher values along the diagonal. Misclassification rates for minority classes are reduced, indicating that the model becomes more capable of distinguishing sentiment classes after balancing. This improvement confirms that data balancing using the N-Gram approach effectively enhances class-level prediction performance, which is consistent with the increases observed in accuracy, precision, recall, and F1-score metrics.

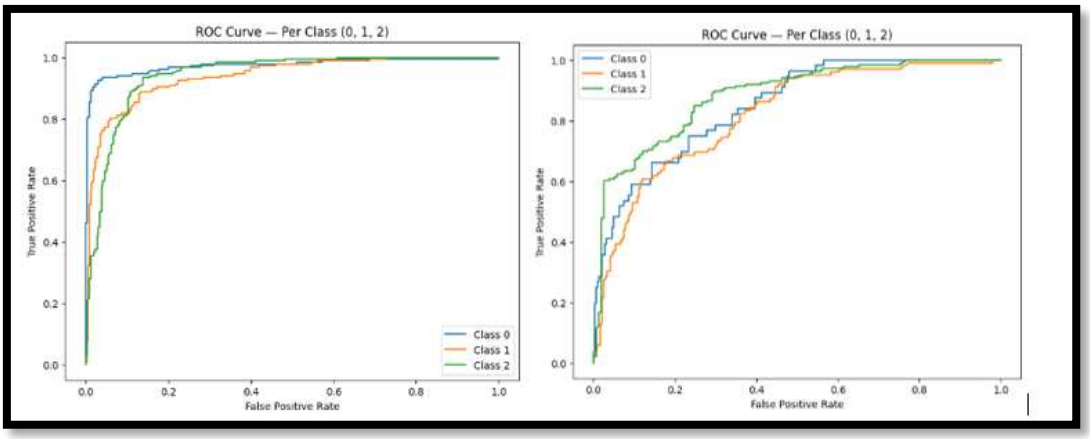


Figure 8. AUC ROC Imbalanced-Balanced

As of shown in figure 8, ROC curves and AUC values are employed to further evaluate the classification performance of the selected best-performing model. The ROC curves for each sentiment class (Class 0, Class 1, and Class 2) on both imbalanced and balanced datasets. Based on the macro-averaged AUC results, the imbalanced dataset achieves a ROC–AUC value of 0.8505, while the balanced dataset shows a substantially higher ROC–AUC value of 0.9526. This improvement indicates that data balancing significantly enhances the model's overall ability to distinguish between sentiment classes. The ROC curves on the balanced dataset demonstrate trajectories closer to the top-left corner across all classes, reflecting higher true positive rates and lower false positive rates compared to the imbalanced dataset. This suggests a more reliable and consistent class discrimination performance after applying the N-Gram-based data balancing approach. These findings are consistent with the improvements observed in accuracy, precision, recall, and F1-score metrics, thereby reinforcing the effectiveness of data balancing in improving classification performance, particularly in multi-class sentiment classification scenarios.

4. CONCLUSION

This study aimed to evaluate the effectiveness of Indonesia Sentiment Lexicon (INSET), address class imbalance using N-Gram-based augmentation and compare machine learning classification performance in Indonesian social media sentiment analysis. The findings show that INSET provides a fast and automated mechanism for initial sentiment labeling through dictionary-based polarity scoring of Indonesian vocabulary. However, its primary limitation lies in its inability to capture contextual meaning, sarcasm, implicit criticism, and domain-specific language, resulting in relatively low labeling accuracy of 55.89% compared with human validation accuracy of 99.70%. Therefore, INSET is suitable for rapid preliminary annotation but requires human validation for high-quality sentiment datasets.

Furthermore, the proposed N-Gram balancing approach successfully addressed class imbalance by generating synthetic minority-class samples while maintaining textual diversity. This significantly improved classification performance across all models, with SVM achieving the best results on balanced data (Accuracy = 0.86, Precision = 0.87, Recall = 0.86, F1-score = 0.86). Overall, the study confirms that combining human-validated labeling and N-Gram-based balancing produces a more reliable sentiment classification framework for Indonesian social media analysis.

REFERENCES

[1] I. Athiyyah Rahma and L. Hulliyyatus Suadaa, "Penerapan Text Augmentation Untuk Mengatasi Data Yang Tidak Seimbang Pada Klasifikasi Teks Berbahasa Indonesia Studi Kasus: Deteksi Judul Clickbait Dan Komentar Hate Speech Pada Berita Online," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1329–1340, 2023, doi: 10.25126/jtiik.2023107325.

[2] C. Engineering, "(Journal of Computer Engineering, System and Science)," vol. 10, no. 1, pp. 136–148, 2025.

[3] M. Abulaish and A. K. Sah, "A Text Data Augmentation Approach for Improving the Performance of CNN," *2019 11th Int. Conf. Commun. Syst. Networks, COMSNETS 2019*, pp. 625–630, 2019, doi: 10.1109/COMSNETS.2019.8711054.

[4] S. A. Nugroho, S. Teknik, I. Fakultas, I. Komputer, and U. K. Soegijapranata, "Comparison Of Support Vector Machine (Svm), Xgboost And Random Forest For Sentiment Analysis Of Bumble App User Comments," vol. 6, no. 1, pp. 32–46, 2022.

[5] R. Hidayat, D. Mahdiana, and A. Fergina, "Comparative Analysis of Logistic Regression , SVM , Xgboost , and Random Forest Algorithms for Diabetes Classification," vol. 7, no. 1, pp. 281–291, 2024, doi: 10.32493/jtsi.v7i1.38258.

[6] N. Epriyanti, A. Meiriza, and D. Y. Hardiyanti, "Perbandingan Kinerja SVM , Random Forest dan

XGBoost pada Aplikasi Access by KAI Menggunakkan ADASYN," vol. 12, no. 5, pp. 733–742, 2025, doi: 10.30865/jurikom.v12i5.9134.

[7] S. Rustad, G. F. Shidik, and I. Nlp, "Ingénierie des Systèmes d ' Information Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus : A Comparative Study of TF-IDF , Word2Vec , FastText BERT , and GPT," vol. 30, no. 6, pp. 1469–1482, 2025.

[8] J. Prasetya, "Leibniz : Jurnal Matematika," vol. 2, pp. 11–22, 2022.

[9] V. Kumar, G. S. Lalotra, P. Sasikala, and D. S. Rajput, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," pp. 1–28, 2022.

[10] G. L. (content published under the G. D. Site), "Datasets: Class-imbalanced datasets," Google for Developers. [Online]. Available: https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

[11] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[12] K. Ramdhan and K. Muslim, "Analisis Sentimen terhadap Toko Online menggunakan Naïve Bayes pada Media Sosial Twitter," *e-Proceeding Eng.*, vol. 5, no. 3, pp. 8141–8151, 2018, [Online]. Available: http://website.com

[13] S. Nanda, D. Mualfah, and D. A. Fitri, "Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest," no. x, pp. 210–219, 2019.

[14] P. P. E. Indarbensyah and N. Rochmawati, "Penerapan N-Gram menggunakan Algoritma Random Forest dan Naïve Bayes Classifier pada Analisis Sentimen Kebijakan PPKM 2021," *J. Informatics Comput. Sci.*, vol. 2, no. 04, pp. 235–244, 2021, doi: 10.26740/jinacs.v2n04.p235-244.

[15] T. Hartina and A. Masri, "Pendeteksi Kesalahan Pengetikan Kata Non Baku pada Karya Tulis Menggunakan Metode N-Gram," *J. Inform.*, vol. 7, no. 1, pp. 77–84, 2020, doi: 10.31311/ji.v7i1.7916.

[16] N. L. Models, "N-gram Language Models," 2025.

[17] M. I. H. A. D. Akbari, A. Novianty, and C. Setianingsih, "Analisis Sentimen Menggunakan Metode Learning Vector Quantization," *e-Proceeding Eng.*, vol. 4, no. 2, p. 2283, 2017, [Online]. Available: https://openlibrary.telkomuniversity.ac.id/pustaka/files/135356/jurnal_eproc/analisis-sentimen-menggunakan-metode-learning-vector-quantization.pdf

[18] R. Guo, Z. Zhao, T. Wang, G. Liu, J. Zhao, and D. Gao, "Degradation state recognition of piston pump based on ICEEMDAN and XGBoost," *Appl. Sci.*, vol. 10, no. 18, pp. 1–17, 2020, doi: 10.3390/APP10186593.

[19] K. Afifah, I. N. Yulita, I. Sarathan, B. Data, and U. Padjadjaran, "Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier," 2021.

[20] C. Science and P. City, "Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors," pp. 27–28, 2021.

[21] J. Informatika, B. N. Setiyono, N. A. Maori, and T. Tamrin, "Analisis Sentimen Ulasan Pengguna Aplikasi Threads di Google Play Menggunakan Algoritma XGBoost Dengan Pen- guatan SMOTE".

[22] M. Y. Khan, A. Qayoom, M. S. Nizami, M. S. Siddiqui, S. Wasi, and S. M. K. Raazi, "Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision , Machine Learning , and Word Embedding-Based Deep Learning Techniques," vol. 2021, 2021.

[23] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM , Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," vol. 16, pp. 141–149, 2022.

[24] M. J. Setiawan, V. Rahmayanti, and S. Nastiti, "DANA App Sentiment Analysis : Comparison of XGBoost , SVM , and Extra Trees," vol. 13, pp. 337–345, 2024.

[25] S. P. Astuti, "Analisis sentimen berbasis aspek pada aplikasi tokopedia menggunakan lda dan naïve bayes," 2020.

[26] R. Dwiyansaputra, S. I. Murpratiwi, and A. Aranta, "Analisis Sentimen Pengguna Platform Media Sosial X Pada Topik Pemilihan Presiden 2024 Menggunakan Perbandingan Model," vol. 9, no. 1, pp. 626–634, 2025.