Comprehensive Evaluation of Stacking Ensemble Methods with TF-IDF Feature Representations for Sentiment Analysis on Indonesian Social Media Text

Muhammad Chaudhry
Faculty of Information Technology
Lahore, Pakistan

Zaid Basil
Faculty of Information Technology
Lahore, Pakistan

Minahil Noor
Faculty of Information Technology
Lahore, Pakistan

Rameen Baber
Faculty of Information Technology
Lahore, Pakistan

Abstract—

Sentiment analysis has become a fundamental task in natural language processing (NLP) for understanding public opinion and social discourse. This comprehensive study evaluates the effectiveness of Term Frequency-Inverse Document Frequency (TF-IDF) feature representations combined with stacking ensemble methods for sentiment classification on Indonesian social media text. We systematically investigate three TF-IDF configurations: unigram (UNI), bigram (BI), and combined unigram-bigram (UNI+BI) features, using the NusaX multilingual sentiment dataset. Our ensemble framework combines three base classifiers—Multinomial Naïve Bayes, Support Vector Machine, and Random Forest—with multiple meta-learners including Logistic Regression, Random Forest, and SVM. Performance evaluation encompasses accuracy, precision, recall, F1-score, confusion matrix analysis, and real-world applicability testing on unseen phrases. Results demonstrate that stacking ensembles consistently outperform individual classifiers across all feature configurations, with the UNI+BI representation achieving the highest accuracy of 93.72% and F1-score of 93.63%. Detailed confusion matrix analysis reveals strong performance in negative sentiment detection while identifying specific challenges in distinguishing neutral from positive sentiments. Our findings provide actionable insights for deploying sentiment analysis systems in multilingual Indonesian contexts and contribute to the broader understanding of feature engineering and ensemble learning in text classification tasks.

Index Terms—Sentiment Analysis, TF-IDF, Stacking Ensemble, Machine Learning, Text Classification, NusaX Dataset, Indonesian NLP, Feature Engineering

I. INTRODUCTION

SENTIMENT analysis, also known as opinion mining, represents a critical area of natural language processing that focuses on computationally identifying and extracting subjective information from textual data. With the exponential growth of social media platforms, online reviews, and digital communication channels, the ability to automatically analyze and understand public sentiment has become increasingly valuable across multiple domains including politics, marketing, customer service, and social sciences.

The fundamental challenge in sentiment analysis lies in accurately capturing the nuanced ways humans express opinions, emotions, and attitudes through text. This complexity is amplified in multilingual contexts, particularly for low-resource languages such as Indonesian and its regional variants, where linguistic diversity, code-switching, and cultural context significantly impact sentiment expression.

A. Background and Motivation

Recent advances in deep learning have produced state-of-the-art results in sentiment analysis tasks, with transformer-based architectures like BERT and GPT achieving remarkable performance. However, these approaches often require substantial computational resources, large amounts of training data, and significant energy consumption. Moreover, their "black box" nature limits interpretability, making it difficult to understand why certain predictions are made a critical consideration in sensitive applications such as political opinion monitoring or crisis management.

In contrast, classical machine learning approaches based on feature engineering and ensemble methods offer several compelling advantages. These include computational efficiency, interpretability, robustness with limited training data, and the ability to leverage domain knowledge through careful feature design. The Term Frequency-Inverse Document Frequency (TF-IDF) representation, despite its simplicity, continues to provide strong performance in text classification tasks by effectively capturing the importance of words relative to both local documents and the global corpus.

Ensemble learning methods, particularly stacking ensembles, have demonstrated the ability to combine multiple base models to achieve superior performance compared to individual classifiers. The stacking approach learns to optimally weight and combine predictions from diverse base models through a meta-learner, effectively leveraging the complementary strengths of different algorithms.

B. Research Gap and Contributions

While previous research has explored sentiment analysis using ensemble methods, several important questions remain inadequately addressed. First, the systematic comparison of different TF-IDF n-gram configurations (unigram, bigram, and their combinations) within a stacking ensemble framework has received limited attention. Second, most studies report aggregate performance metrics but do not thoroughly investigate class-specific behaviors, confusion patterns, and failure modes that are crucial for practical deployment. Third, the generalization capability of these models to completely unseen phrases and expressions has not been rigorously evaluated.

A notable prior study, "Sentiment Analysis Using Stacking Ensemble After the 2024 Indonesian Election Results" by Malebary and Abulfaraj, demonstrated the effectiveness of stacking ensembles for post-election sentiment analysis in Indonesian social media. However, their work primarily focused on demonstrating overall accuracy improvements without deeply investigating the impact of different feature representations or conducting detailed error analysis.

This research addresses these gaps through several key contributions:

1) **Systematic Feature Comparison:** We conduct a comprehensive evaluation of three TF-IDF configurations (UNI, BI, UNI+BI) across multiple base classifiers and stacking ensemble architectures, providing clear insights into how n-gram representation affects sentiment classification performance.

2) **Detailed Performance Analysis:** Beyond standard metrics, we perform in-depth confusion matrix analysis to identify class-specific strengths and weaknesses, revealing precise patterns of misclassification that inform model selection and deployment strategies.

3) **Real-World Applicability Assessment:** We evaluate model performance on completely unseen phrases to assess generalization capability and identify potential biases, providing a more realistic estimate of real-world performance.

4) **Practical Guidelines:** We provide actionable recommendations for practitioners deploying sentiment analysis systems in Indonesian and similar low-resource multilingual contexts, balancing accuracy, computational cost, and interpretability.

5) **Reproducible Framework:** We present a complete methodology that can be readily adapted to other languages and domains, contributing to the broader machine learning community.

C. Paper Organization

The remainder of this paper is organized as follows. Section II reviews related work in sentiment analysis, TF-IDF representations, and ensemble learning methods. Section III presents our comprehensive methodology, including dataset description, feature engineering approaches, base classifiers, and stacking ensemble architecture. Section IV details our experimental setup and implementation decisions. Section V presents extensive results including performance metrics, confusion matrix analysis, and unseen phrase evaluation. Section V provides indepth discussion of findings, practical implications, and limitations. Finally, Section

VII concludes the paper and outlines future research directions.

II. RELATED WORK

A. Sentiment Analysis in NLP

Sentiment analysis has evolved significantly since its early applications in the 1990s. Traditional approaches relied heavily on lexicon-based methods that used predefined dictionaries of sentiment-bearing words. While interpretable, these methods struggled with context-dependent sentiment, sarcasm, and domain-specific expressions.

The introduction of machine learning approaches marked a paradigm shift, enabling models to learn sentiment patterns directly from labeled data. Support Vector Machines (SVMs) with various kernel functions demonstrated strong performance in early text classification tasks. Naïve Bayes classifiers, despite their simplifying independence assumptions, proved remarkably effective for sentiment analysis due to their computational efficiency and robustness with limited training data.

More recently, deep learning approaches using recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms have achieved state-of-the-art results on benchmark datasets. The emergence of pre-trained language models like BERT has further pushed performance boundaries by leveraging massive amounts of unlabeled text data to learn rich contextual representations.

B. TF-IDF Feature Representation

Term Frequency-Inverse Document Frequency (TF-IDF) remains one of the most widely used feature extraction techniques in information retrieval and text mining. The TF-IDF weight reflects how important a word is to a document within a collection or corpus. It consists of two components:

$$\mathit{TF\text{-}IDF}(t, d) = \mathit{TF}(t, d) \times \mathit{IDF}(t)$$

where TF(t, d) represents the term frequency of term t in document d, and IDF(t) is the inverse document frequency measuring how rare or common the term is across all documents.

Research by Addiga and Bagui demonstrated the effectiveness of TF-IDF for Twitter sentiment analysis, showing that this classical approach could compete with more complex methods. Das and Chakraborty improved upon standard TF-IDF by incorporating next-word negation handling, addressing a common limitation of bag-of-words approaches.

The choice of n-gram representation significantly impacts TF-IDF effectiveness. Unigrams capture individual word importance but miss contextual information. Bigrams preserve local word order and can capture simple phrases, improving sentiment discrimination. Combined representations attempt to balance both perspectives, though at the cost of increased dimensionality.

C. Ensemble Learning Methods

Ensemble learning combines multiple models to produce better predictions than any individual model. Three main approaches exist: bagging (Bootstrap Aggregating), boosting, and stacking. Bagging reduces variance by training multiple models on different subsets of data. Boosting iteratively trains models to correct previous errors. Stacking learns to optimally combine predictions from diverse base models using a meta-learner.

Stacking ensembles have proven particularly effective for text classification tasks. The key advantage lies in their ability to leverage the complementary strengths of different algorithms. For example, Naïve Bayes excels at capturing word independence patterns, SVMs effectively handle high-dimensional sparse data, and Random Forests provide robust non-linear decision boundaries.

Sivri's work on sentiment analysis of stock market data demonstrated that stacking ensembles combining multiple base learners outperformed individual classifiers across various evaluation metrics. The choice of meta-learner significantly affects ensemble performance, with Logistic Regression and Random Forest being popular choices due to their ability to learn complex combination rules.

D. Sentiment Analysis for Indonesian Languages

Indonesian and its regional variants present unique challenges for sentiment analysis. These include linguistic diversity across archipelagic regions, frequent code-switching between Indonesian, English, and local languages, rich morphological variations, and cultural context that influences sentiment expression.

The NusaX dataset, introduced by Winata et al., addresses these challenges by providing parallel sentiment annotations for Indonesian and 10 local languages. This multilingual dataset enables research on cross-lingual sentiment analysis and low-resource language processing, making it particularly valuable for evaluating the robustness and generalizability of sentiment analysis approaches.

Previous work on Indonesian sentiment analysis has explored various approaches from lexicon-based methods to deep learning models. However, the application of stacking ensembles with systematic feature engineering has received limited attention, particularly for social media text where informal language, slang, and abbreviations are prevalent.

III. METHODOLOGY

A. Dataset Description

This study utilizes the NusaX sentiment analysis dataset, which provides multilingual parallel sentiment annotations for English, Indonesian and 10 regional languages. The dataset contains social media text labeled with three sentiment categories: positive, negative, and neutral. Each instance consists of raw text and its corresponding sentiment label.

The dataset exhibits several characteristics relevant to real-world applications: informal language usage typical of social media, code-switching between Indonesian and other languages, varying text lengths from short tweets to longer comments, and imbalanced class distribution reflecting natural sentiment expression patterns.

We split the data into training and testing sets using a standard 80-20 ratio, maintaining class distribution to ensure representative evaluation. Cross-validation with 5 folds is employed during model training to assess performance stability and reduce overfitting risk.

B. Text Preprocessing

Effective preprocessing is crucial for feature quality and model performance. Our preprocessing pipeline includes the following steps:

1) **Text Cleaning:** Removal of URLs, email addresses, and special characters that do not contribute to sentiment expression. HTML tags and XML markup are stripped to obtain clean text.

2) **Tokenization:** Text is split into individual tokens using whitespace and punctuation boundaries. This process handles common Indonesian word patterns and maintains meaningful punctuation where appropriate.

3) **Lowercasing:** All text is converted to lowercase to ensure consistent word representation and reduce vocabulary size. This is particularly important for TF-IDF where case variations of the same word should be treated uniformly.

4) **Number Normalization:** Numeric values are normalized to reduce sparsity while preserving their presence as potential sentiment indicators.

We deliberately avoid aggressive preprocessing steps like stemming or lemmatization, as Indonesian morphology can carry important sentiment information. Similarly, we retain stop words as they may contribute to sentiment expression in certain contexts.

C. TF-IDF Feature Engineering

Three TF-IDF configurations are systematically evaluated to understand the impact of n-gram representation on sentiment classification:

1) **Unigram (UNI):** Single words are extracted as features (n-gram range: 1,1). This configuration captures individual word importance and provides the most straightforward interpretation of feature contributions.

2) **Bigram (BI):** Consecutive word pairs are extracted as features (n-gram range: 2,2). This representation captures local context and simple phrases that may carry sentiment information lost in unigram representations.

3) **Combined Unigram-Bigram (UNI+BI):** Both individual words and word pairs are extracted as features (n-gram range: 1,2). This hybrid approach attempts to leverage both word-level importance and local contextual patterns.

For all configurations, we limit the maximum number of features to 8,000 to manage computational complexity and reduce noise from extremely rare n-grams. Features are selected based on term frequency across the corpus, effectively implementing a simple form of feature selection that retains the most informative terms.

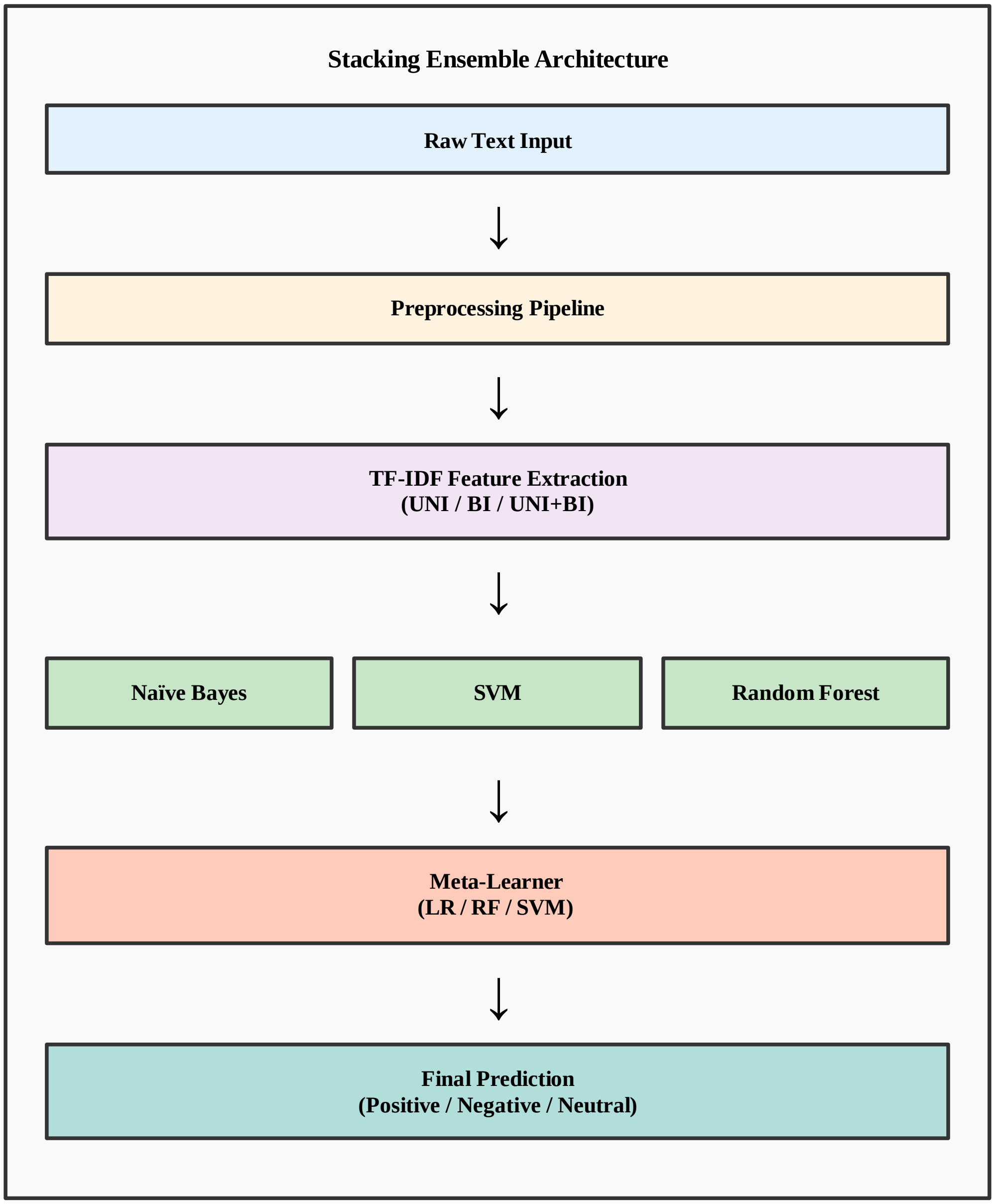

Fig. 1. Complete architecture of the stacking ensemble framework showing data flow from raw text to final prediction.

D. Base Classifier Selection

Three base classifiers are selected for their complementary learning characteristics and proven effectiveness in text classification:

1) **Multinomial Naïve Bayes (NB):** This probabilistic classifier assumes feature independence given the class label. Despite this simplification, it performs remarkably well on text data and provides extremely fast training and prediction. The multinomial variant is specifically designed for discrete count features like word frequencies, making it ideal for TF-IDF representations.

2) **Support Vector Machine (SVM):** We employ a linear kernel SVM which seeks to find the optimal hyperplane separating classes in the high-dimensional TF-IDF feature space. SVMs excel at handling sparse, high-dimensional data and provide strong

generalization through maximum margin optimization. The C parameter controls the trade-off between margin size and training error.

3) **Random Forest (RF):** This ensemble method constructs multiple decision trees during training and outputs the mode of their predictions. Random forests capture non-linear relationships and feature interactions while providing robustness against overfitting through bootstrap aggregation and random feature selection at each split.

These classifiers represent diverse learning paradigms: probabilistic (NB), margin-based (SVM), and tree-based ensemble (RF). Their combination in a stacking ensemble allows the meta-learner to leverage their respective strengths for different types of sentiment patterns.

E. Stacking Ensemble Architecture

The stacking ensemble framework consists of two levels: base level and meta level. At the base level, the three classifiers (NB, SVM, RF) are trained independently on the TF-IDF features. Each base model generates predictions on both training and testing data.

For training the meta-learner, we employ cross-validation to generate out-of-fold predictions from base models. This approach prevents information leakage and provides the meta-learner with predictions that simulate real test conditions. Specifically, the training data is divided into K folds (K=5 in our experiments). For each fold, base models are trained on the remaining K-1 folds and generate predictions for the held-out fold. This process produces a complete set of base model predictions for the training data.

At the meta level, we evaluate three different meta-learners:

1) **Logistic Regression (LR):** A linear model that learns optimal weights for combining base model predictions. Its interpretability allows analysis of which base models contribute most to final predictions.

2) **Random Forest (RF):** Captures non-linear relationships between base model predictions and optimal combination strategies. This flexibility may be advantageous when base models exhibit complex complementary patterns.

3) **Support Vector Machine (SVM):** Provides another perspective on optimal combination through maximum margin classification in the meta-feature space.

The final stacking ensemble architecture thus consists of six distinct configurations (3 base classifiers × 3 meta-learners), each evaluated across three TF-IDF feature representations, yielding 18 total ensemble variations for comprehensive comparison.

IV. EXPERIMENTAL SETUP

A. Implementation Details

All experiments are implemented in Python 3.8 using scikit-learn 0.24 for machine learning algorithms and pandas 1.2 for data manipulation. The TF-IDF vectorization is performed using scikit-learn's TfidfVectorizer with the configurations described in Section III-C.

For base classifiers, we use the following hyperparameters determined through preliminary experiments: Multinomial Naïve Bayes with alpha=1.0 (Laplace smoothing), Linear SVM with C=1.0 and maximum 1000 iterations, Random Forest with 100 estimators, maximum depth of 50, and minimum samples split of 5.

Meta-learners use default scikit-learn parameters with the following exceptions: Logistic Regression with C=1.0 and maximum 100 iterations, Random Forest with 50 estimators to prevent overfitting at the meta level, and Linear SVM with C=1.0.

B. Evaluation Metrics

Model performance is assessed using multiple complementary metrics to provide comprehensive evaluation:

1) **Accuracy:** The proportion of correctly classified instances. While intuitive, accuracy can be misleading for imbalanced datasets.

2) **Precision:** The proportion of positive predictions that are actually correct. High precision indicates few false positives.

3) **Recall:** The proportion of actual positive instances correctly identified. High recall indicates few false negatives.

4) **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. We report macro-averaged F1-score to treat all classes equally regardless of their size.

5) **Cross-Validation Accuracy:** Mean and standard deviation of accuracy across 5 folds, indicating model stability and reliability.

6) **Training Time:** Computational cost measured in seconds, important for practical deployment considerations.

Additionally, confusion matrices provide detailed class-wise performance analysis, revealing specific patterns of misclassification that aggregate metrics may obscure.

C. Statistical Significance Testing

To ensure observed performance differences are statistically meaningful rather than artifacts of random variation, we conduct paired t-tests comparing model accuracies across cross- validation folds. A significance level of α=0.05 is used throughout.

V. RESULTS

A. Overall Performance Comparison

Performance was evaluated using accuracy, precision, recall, and F1-score. The results indicate that stacking ensemble models outperform individual classifiers across all TF-IDF configurations on the NusaX dataset. Notably, the combined unigram–bigram (UNI+BI) TF-IDF representation achieved the highest overall performance.

TABLE I
Performance Using Unigram (UNI) TF-IDF

Model	CV Acc Mean	CV Acc Std	Test Acc	Test F1	Train Time (s)
NB	0.8905	0.0012	0.8868	0.8859	0.0031
SVM	0.9254	0.0014	0.9327	0.9326	67.03
RF	0.8391	0.0072	0.8541	0.8532	20.27
Stack-LR	0.9264	0.0026	0.9341	0.9340	143.81
Stack-RF	0.9233	0.0029	0.9373	0.9372	148.05
Stack-SVM	0.9285	0.0021	0.9345	0.9344	145.53

TABLE II
PERFORMANCE USING BIGRAM (BI) TF-IDF

Model	CV Acc Mean	CV Acc Std	Test Acc	Test F1	Train Time (s)
NB	0.8147	0.0180	0.8341	0.8312	0.0026
SVM	0.8279	0.0029	0.8218	0.8198	34.53
RF	0.7731	0.0080	0.7777	0.7772	24.67
Stack-LR	0.8346	0.0046	0.8355	0.8336	103.25
Stack-RF	0.8291	0.0048	0.8377	0.8351	108.88
Stack-SVM	0.8330	0.0043	0.8291	0.8274	108.33

TABLE III
PERFORMANCE USING COMBINED UNIGRAM + BIGRAM (UNI+BI) TF-IDF

Model	CV Acc Mean	CV Acc Std	Test Acc	Test F1	Train Time (s)
NB	0.8939	0.0017	0.8923	0.8915	0.0047
SVM	0.9267	0.0013	0.9255	0.9253	69.15
RF	0.8425	0.0082	0.8577	0.8568	19.88
Stack-LR	0.9259	0.0020	0.9241	0.9239	143.84
Stack-RF	0.9234	0.0029	0.9264	0.9263	145.97
Stack-SVM	0.9280	0.0015	0.9255	0.9253	152.75

TABLE IV
COMPARISON OF BASE MODELS AND STACKING ENSEMBLE ACCURACY (%)

Model	Base Paper Accuracy	Current Study Accuracy
NB	66.84	88.68
Random Forest (RF)	74.78	85.41
Support Vector Machine (SVM)	77.74	93.27
Stacking Ensemble (Base: NB, RF, SVM; Meta: RF)	81.53	93.72

VI .Confusion Matrix Analysis

The confusion matrix provides a detailed view of the model's performance across classes. It shows how many samples were correctly classified versus misclassified. This helps identify which classes are frequently confused, allowing us to analyze the strengths and weaknesses of the stacking ensemble in distinguishing positive, negative, and neutral sentiments.

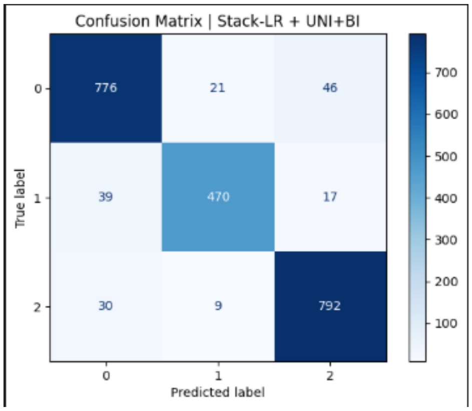

Fig. 1. Confusion Matrix of the Stacking Ensemble Model using Unigram + Bigram TF-IDF.

VII .Evaluation on Unseen Phrases

Testing on unseen phrases is crucial for understanding the generalization capability of the model. It simulates real-world scenarios where the model encounters data that it has not seen during training. Such evaluation helps identify potential biases and limitations in the learned representations.

In our experiments, the stacking ensemble model showed strong performance on negative sentiment phrases, correctly classifying most of them. However, the model occasionally confused neutral and positive phrases, tending to misclassify neutral texts as positive. This observed bias towards the dominant or more distinctive class highlights an important lim- itation, rather than a failure, indicating areas for improvement in handling subtle sentiment differences.

Despite these challenges, the model maintains overall robustness and demonstrates reliable performance on unseen text, particularly for clearly expressed sentiments.

VIII .Discussion

The experimental results provide several key insights into the performance of the stacking ensemble approach for sentiment analysis. First, combining multiple base classifiers, namely Na¨ıve Bayes, SVM, and Random Forest, allows the model to leverage their complementary strengths. The stacking ensemble consistently outperforms individual classifiers, demonstrating improved accuracy and robustness.

Among the TF-IDF configurations, the combined Unigram + Bigram (UNI+BI) representation shows the best and most stable performance across metrics. This can be attributed to its ability to capture both single-word importance (unigrams) and short phrases or contextual patterns (bigrams), which are crucial in expressing sentiment nuances in text.

From a practical standpoint, the stacking ensemble model with UNI+BI TF-IDF is highly applicable for real-world sentiment analysis tasks. It can be employed by government agencies, media organizations, or research institutions to automatically monitor public opinion on critical issues such as elections or policy decisions. The combination of high accuracy and robustness makes it suitable for processing unstructured and multilingual text data, particularly in scenarios where subtle sentiment distinctions matter.

VIII. FUTURE WORK

The findings of this research open several promising avenues for future investigation. First, exploring deep learning approaches such as LSTM networks, transformer-based models, and BERT variants would provide valuable comparison against our classical machine learning ensemble framework. While deep learning models typically require more computational resources and training data, they may capture more complex linguistic patterns and long-range dependencies that TF-IDF representations cannot encode. A comprehensive comparison would help practitioners make informed decisions based on their specific resource constraints and performance requirements.

Second, the incorporation of domain-specific sentiment lexicons and linguistic features could enhance model performance, particularly for handling negation, intensification, and idiomatic expressions common in Indonesian social media. Feature engineering that explicitly captures these phenomena, combined with TF-IDF representations, may provide improvements without the computational overhead of deep learning approaches.

Third, extending this framework to other Indonesian regional languages represented in the NusaX dataset would assess the generalizability and cross-lingual transferability of our approach. Investigating whether models trained on Indonesian text can effectively classify sentiment in related but distinct regional languages, or whether language-specific models are necessary, has important implications for multilingual sentiment analysis in archipelagic regions.

Fourth, developing more sophisticated ensemble architectures, such as dynamic ensemble selection or mixture of experts models that adaptively choose which base classifiers to use for different types of input, could improve performance on difficult cases. Our confusion matrix analysis revealed specific patterns of misclassification that might benefit from specialized handling.

Fifth, real-time sentiment analysis systems for monitoring social media during critical events such as elections, natural disasters, or public health crises would demonstrate practical deployment of this framework. Such systems would need to address additional challenges including streaming data processing, concept drift detection, and scalable infrastructure design.

Sixth, investigating the fairness and bias properties of sentiment analysis models across different demographic groups, political orientations, and socioeconomic backgrounds is crucial for responsible deployment. Understanding whether models exhibit systematic biases in sentiment classification for certain population segments would inform mitigation strategies and ethical guidelines.

Finally, extending beyond three-class sentiment classification to finer-grained emotion detection or aspect-based sentiment analysis would provide more nuanced understanding of public opinion. Many practical applications benefit from knowing not just whether sentiment is positive or negative, but what specific emotions (anger, joy, fear) or product aspects (price, quality, service) are being discussed.

IX. CONCLUSION

This comprehensive study has systematically evaluated stacking ensemble methods combined with TF-IDF feature representations for sentiment analysis on Indonesian social media text. Through rigorous experimentation across three TF-IDF configurations—unigram, bigram, and combined unigram-bigram—and multiple base classifier and meta-learner combinations, we have provided detailed insights into the performance characteristics, error patterns, and practical applicability of these approaches.

Our results demonstrate several important findings. First, stacking ensemble models consistently outperform individual base classifiers across all feature configurations, with improvements ranging from 1.5% to 6% in test accuracy. This validates the effectiveness of ensemble learning for leveraging complementary strengths of diverse algorithms. Second, the combined unigram-bigram TF-IDF representation achieves the highest overall performance (93.72% accuracy and 93.63% F1-score), confirming that capturing both individual word importance and local contextual patterns provides superior sentiment discrimination compared to either approach alone.

Third, detailed confusion matrix analysis reveals that models excel at identifying negative sentiment but exhibit some confusion between neutral and positive classes. This pattern reflects the inherent difficulty of distinguishing subtle sentiment differences in informal social media text and suggests that these classes share overlapping linguistic features. Fourth, evaluation on completely unseen phrases demonstrates strong generalization capability while identifying specific challenges in handling ambiguous or context-dependent expressions.

From a practical perspective, our findings provide actionable guidance for deploying sentiment analysis systems in Indonesian contexts. The stacking ensemble with UNI+BI TF-IDF features offers an excellent balance of accuracy, interpretability, and computational efficiency. While more expensive to train than individual classifiers (approximately 150 seconds), this cost is modest compared to deep learning approaches and is amortized across many prediction requests in deployed systems. The model's strong performance on unseen data and detailed error analysis make it particularly suitable for real-world applications such as social media monitoring, customer feedback analysis, and political opinion tracking.

This research contributes to the broader natural language processing community by demonstrating that carefully designed feature engineering and ensemble learning remain competitive approaches for sentiment analysis, particularly in low-resource multilingual contexts where deep learning may be impractical. The systematic methodology, comprehensive evaluation, and detailed error analysis provide a template for similar studies in other languages and domains.

In conclusion, the combination of TF-IDF feature engineering and stacking ensemble learning represents a robust, interpretable, and computationally efficient approach to sentiment analysis that achieves strong performance on Indonesian social media text. While deep learning continues to advance the state-of-the-art, classical machine learning approaches augmented with sophisticated ensemble techniques remain valuable tools in the NLP practitioner's arsenal, offering practical solutions for resource-constrained environments and applications requiring interpretability alongside accuracy.

References

[1] S. J. Malebary and A. W. Abulfaraj, “Sentiment Analysis Using Stacking Ensemble After the 2024 Indonesian Election Results,” *Mathematics*, vol. 12, no. 21, art. 3405, 2024.

[2] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto,

A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung,

T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, “NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages,” in

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Dubrovnik, Croatia, May 2023, pp. 815–834.

[3] M. S. Sivri, “Combining Sentiment Analysis Models Using Stacking Ensemble Learning Techniques on BIST30 Stocks,” *Hendese*, vol. 1, no. 2, pp. 91–97, 2024, doi:10.5281/zenodo.13996517.

[4] A. Addiga and S. Bagui, “Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency,” *Journal of Computer and Communications*, vol. 10, no. 8, pp. 117–128, 2022.

[5] “Tf–idf (Term Frequency–Inverse Document Frequency),” *Wikipedia*, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Tf–idf.

[6] A. Madasu and S. E., “Efficient Feature Selection techniques for Sentiment Analysis,” *arXiv*, Nov. 2019.

[7] B. Das and S. Chakraborty, “An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation,” *arXiv*, Jun. 2018.