Jurnal Masyarakat Informatika, 16(1), 2025
DOI:10.14710/jmasif.16.1.73115
ISSN: 2777-0648

An Ensemble-Based Approach for Detecting Clickbait in Indonesian Online Media

Sandy Kurniawan*, Adhe Setya Pramayoga, and Yeva Fadhilah Ashari
Department of Informatics, Universitas Diponegoro
Semarang, Indonesia
Corresponding author: sandy@live.undip.ac.id

Abstract
Clickbait headlines are widely used in online media to attract readers through exaggerated or misleading titles, potentially leading to user dissatisfaction and information overload. This study proposes a machine learning approach for detecting clickbait in Indonesian news headlines using classical classification models and ensemble learning. The dataset consists of labeled clickbait and non-clickbait headlines in Bahasa Indonesia, which were processed and represented using TF-IDF vectorization. Two preprocessing scenarios, with and without stopwords removal, were explored to examine the impact of common but often semantically irrelevant words on classification performance. Three base classifiers, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine, were integrated using soft voting and stacking ensemble methods. The experimental results indicate that the stacking ensemble model achieved the highest accuracy of 0.7728, while the voting ensemble recorded the best F1-score of 0.7080, outperforming individual classifiers. Despite these gains, the SVM model demonstrated the most substantial decline in accuracy after stopwords removal, dropping by 0.0410. These findings highlight the effectiveness of ensemble learning in enhancing clickbait detection and the importance of preprocessing choices in model performance.

Keywords : *Clickbait Detection, Ensemble Learning, Text Classification, Stacking, Voting, Stopwords*

1 Introduction

In today's digital era, the proliferation of online news and content has significantly altered how information is consumed. Headlines have become a critical component in capturing user attention, often determining whether an article is read or ignored. In response to this, many online publishers employ clickbait—headlines that are intentionally sensationalized, misleading, or exaggerated to entice readers to click [1], [2]. While clickbait can be effective in driving traffic, it often leads to user dissatisfaction, misinformation, and diminished trust in media sources [3], [4].

The detection of clickbait has thus emerged as an important task in natural language processing (NLP) and media quality control. While prior studies have explored deep learning models for this task [5], such approaches can be computationally expensive and difficult to interpret. In contrast, traditional machine learning models, when paired with proper text preprocessing and feature extraction techniques, can still offer competitive performance while being lightweight and interpretable [6], [7].

This study investigates the effectiveness of classical machine learning models, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machines, in detecting clickbait headlines written in Indonesian. The research includes a detailed comparison of models trained on two

Received: 10 May 2025 ; Revised: 25 May 2025; Accepted: 30 May 2025; Published: 31 May 2025

preprocessing variants: one with stopwords retained and another with stopwords removed. Furthermore, ensemble learning techniques such as soft voting and stacking classifiers are applied to enhance prediction robustness. The main contribution of this work is a systematic experimental evaluation of various traditional machine learning models and ensemble strategies for Indonesian clickbait headline detection. In particular, the study explores the impact of stopword removal on classification performance and highlights the practicality of combining simple yet effective algorithms through ensemble learning. The findings offer insights into building efficient and interpretable clickbait detection systems tailored to the Indonesian language context.

2 Literature Review

Clickbait is a headline writing strategy designed to attract internet users to click on a hyperlink, typically by highlighting sensational, provocative, or even misleading elements. This phenomenon has raised concerns due to its potential to degrade the quality of information and diminish public trust in online media [8]. Clickbait detection is a task aimed at identifying and filtering web content specifically designed to generate clicks. To address this issue, various approaches have been developed for clickbait detection, which is generally framed as a binary classification task distinguishing between clickbait and non-clickbait content based on information such as article headlines or body text [9].

Clickbait detection has been approached using both traditional machine learning and deep learning techniques. Classical machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), and Random Forest have been widely applied in this domain [10], [11], [12]. Meanwhile, deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer-based architectures have often been employed to enhance detection performance by capturing complex patterns within the data [12], [13]. Although these methods have demonstrated promising results, the majority of studies remain focused on English-language datasets.

To bridge this gap, several recent studies have begun to explore clickbait detection in Indonesian news articles [14], [15]. One of the relevant datasets for this purpose is CLICK-ID, which consists of Indonesian-language news headlines labeled as either clickbait or non-clickbait [5]. This study leverages the CLICK-ID dataset to implement an ensemble learning approach by combining multiple traditional machine learning algorithms, including Naïve Bayes, Logistic Regression, and Support Vector Machine. By applying both voting and stacking classifier strategies, this research aims to assess the extent to which model combinations can improve clickbait classification performance in the Indonesian language context, which remains relatively underexplored.

3 Research Methods

This study employs a structured experimental approach comprising six key stages: dataset collection, text preprocessing, feature extraction, model development, ensemble learning, and evaluation. Each phase is designed to support the main objective of the research: to classify Indonesian news headlines as clickbait or non-clickbait using conventional machine learning algorithms and ensemble methods. The general workflow of this research is illustrated in Figure 1.

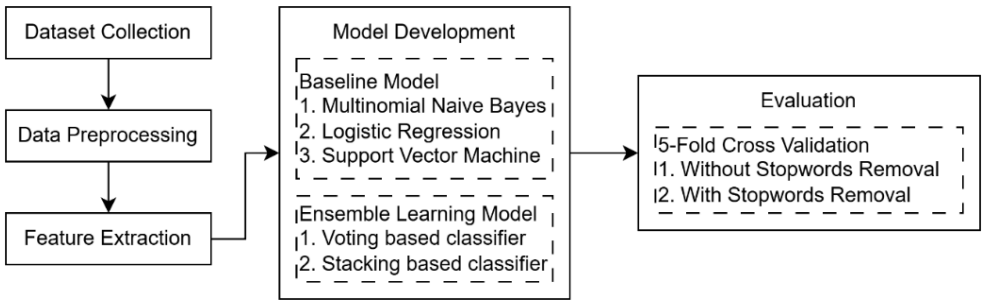

Figure 1 Clickbait Detection General Workflow

3.1 Dataset Collection

This study utilized the publicly available CLICK-ID dataset, which was obtained from a prior research initiative focusing on Indonesian-language clickbait detection [5]. The dataset, formatted in JSON, contains news headlines accompanied by categorical labels indicating whether the headline is a clickbait or non-clickbait. For this study, the categorical labels were mapped into binary values, where "clickbait" was encoded as 1 and "non-clickbait" as 0. The data was loaded into a Pandas DataFrame, enabling further processing and exploration.

To better understand the characteristics of the dataset prior to model development, a descriptive statistical analysis was conducted. This analysis includes label distribution, headline length, and the effects of preprocessing steps such as stopword removal. These insights are essential to assess data balance and potential biases that could influence classification performance. A summary of the dataset statistics is presented in Table 1, Table 2, and Table 3, while the visual distribution of the labels is illustrated in Figure 2 and Figure 3. Figure 2 shows the proportion of clickbait and non-clickbait labels, while Figure 3 visualizes the distribution of headline lengths

Table 1 Descriptive Statistics of the Clickbait Dataset

No.	Statistic	Description
1	Total number of data	15000
2	Number of clickbait data	6290
3	Number of non-clickbait data	8710
4	Proportion of clickbait	41.93%
5	Proportion of non-clickbait	58.07%

Table 2 Title Length Statistics (Before Preprocessing)

No.	Statistic	Character Length	Word Count
1	Minimum	12	2
2	Maximum	123	19
3	Average	64.28	9.68
4	Median (50%)	63	9
5	Standard Deviation (std)	14.32	2.33

Table 3 Average Title Length by Label

Label	Average Characters	Average Words
Clickbait	68.29	10.35
Non-Clickbait	61.39	9.2

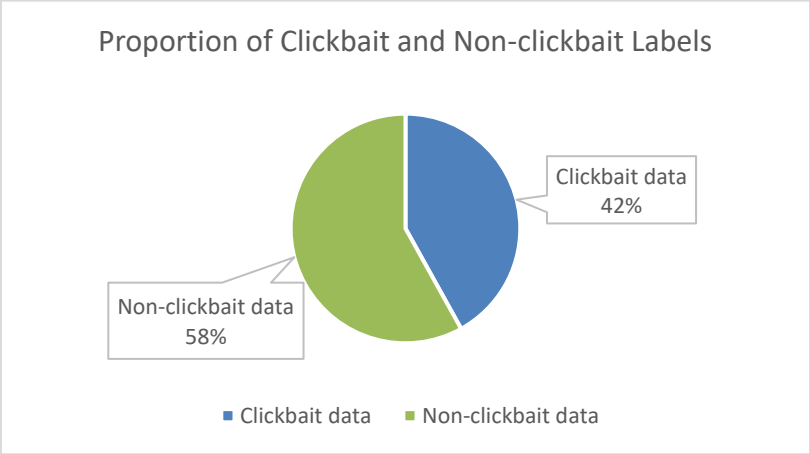

Figure 2 Proportion of Clickbait and Non-clickbait Headlines in the Dataset.

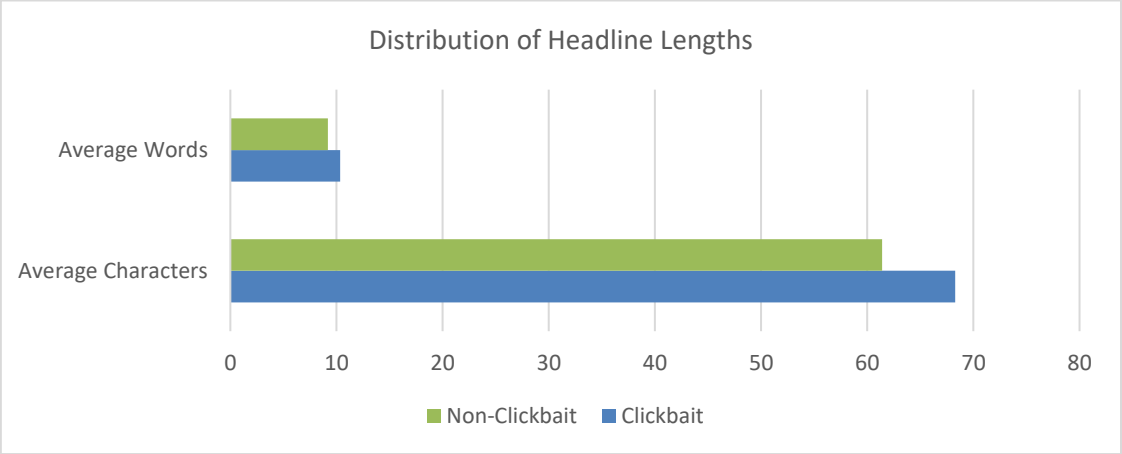

Figure 3 Average Headline Length by Label in Characters and Words

3.2 Text Preprocessing

To prepare the textual data for modeling, several preprocessing steps were performed. Initially, all headline texts were lowercased, and non-alphanumeric characters were removed [16]. Tokenization was applied using NLTK's word tokenizer, followed by stemming using the Sastrawi stemmer, which is specifically built for the Indonesian language. Two parallel versions of the processed data were created: one that retained stopwords and another where stopwords were removed using NLTK's Indonesian stopword list. This dual-track preprocessing was designed to analyze the effect of stopword presence on classification accuracy.

3.3 Feature Extraction

After the preprocessing stage, each sentence is transformed into a numerical vector using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This results in a feature matrix that represents the weighted importance of terms based on their relative frequency within individual documents and across the entire corpus. To capture contextual information, both unigrams and bigrams are utilized as features, with a maximum of 5,000 features selected to reduce the risk of overfitting. The resulting feature matrix serves as input for the classification algorithms in the subsequent modeling phase.

3.4 Model Development

In the model development stage, this study employs three classical machine learning algorithms as baseline models: Multinomial Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). Multinomial Naïve Bayes was selected due to its suitability for text data represented as word frequencies or term frequency-inverse document frequency (TF-IDF). Its probabilistic nature also enables efficient training and testing, especially when dealing with high-dimensional feature spaces [17], [18]. Logistic Regression was chosen for its capability to directly address binary classification problems. Moreover, it is a flexible and well-established algorithm that is often used as a standard benchmark in text classification experiments, making it a reliable baseline [19]. Support Vector Machine with a linear kernel was selected for its effectiveness in handling high-dimensional text data. In the context of text classification, SVM is well-known for its ability to process sparse feature matrices and its robustness against overfitting, particularly when training data is limited [20], [21].

3.5 Ensemble Learning

This study also implements ensemble learning techniques to explore potential performance enhancements in Indonesian-language clickbait detection. Two ensemble strategies are employed: a voting classifier and a stacking classifier. The voting classifier aggregates the output probabilities from three baseline models—Multinomial Naïve Bayes, Logistic Regression, and Support Vector Machine—using soft voting, in which the final class label is determined by averaging the class probabilities predicted by each model. This method was selected for its simplicity and effectiveness in combining diverse classifiers [22], [23]. Meanwhile, the stacking approach uses the same three models as base learners, with Logistic Regression serving as the meta-learner. This method aims to improve predictive accuracy by combining both the original input features and the base models' outputs, thus capturing more complex relationships in the data [24], [25].

3.6 Evaluation

To assess the performance of the proposed models in detecting clickbait headlines, we employed several standard evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the classification effectiveness, particularly in handling the binary nature of the clickbait detection task. To ensure the reliability and generalizability of the evaluation, we adopted a 5-fold cross-validation approach. In this method, the dataset is divided into five equal subsets. Each subset is used once as the validation set while the remaining four subsets serve as the training set. This process is repeated five times, and the final evaluation results are obtained by averaging the performance across all folds. By doing this, we mitigate the risk of overfitting and reduce the variance caused by random partitioning of the data.

Furthermore, we compared the performance of individual baseline models (Multinomial Naïve Bayes, Logistic Regression, and Support Vector Machine) against the proposed ensemble learning approaches, including both voting and stacking classifiers. The purpose of this comparison is to determine whether the ensemble methods offer significant improvements over the standalone classifiers in terms of accuracy and robustness. Through this evaluation procedure, we aim to validate the effectiveness of ensemble learning for clickbait detection in the Indonesian.

4 Results and Discussion

This section describes the experimental scenario and the analysis of the clickbait detection task on Indonesian news headlines. The discussion centers on assessing the impact of stopwords preprocessing on model performance, comparing the effectiveness of two approaches: without stopwords removal and with stopwords removal.

4.1 Experiment Scenario

To assess the effectiveness of the proposed method in detecting clickbait in Indonesian news headlines, a series of experiments were conducted using classical machine learning and ensemble learning approaches. The experiments were structured to compare the impact of two preprocessing strategies: with and without stopwords removal. This comparison was intended to evaluate the influence of stopwords on the predictive performance of the models. The remaining preprocessing steps were kept consistent, including tokenization, stemming, and transformation into feature vectors using TF-IDF with unigram and bigram representations. The number of features was capped at 5000 to mitigate the risk of overfitting and to optimize computational efficiency.

Three classical machine learning algorithms were employed as baseline models: Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine. In addition, two ensemble learning strategies were implemented:

1. Voting Classifier, utilizing soft voting to aggregate predictions from the three baseline models.
2. Stacking Classifier, which incorporates the same three base models with Logistic Regression as the meta-learner. The passthrough attribute was activated to allow the original data features to be included in the stacking process, potentially enhancing predictive accuracy.

Each model was evaluated using a 5-fold stratified cross-validation to ensure balanced class distributions across folds. Model performance was subsequently assessed using accuracy, precision, recall, and F1-score.

4.2 Results and Discussion

This section presents the evaluation results of the proposed methods for detecting clickbait in Indonesian news headlines. The analysis is divided into two parts: without stopwords removal and with stopwords removal. Additionally, the impact of ensemble learning methods on model performance is examined, focusing on accuracy, precision, recall, and F1-score.

4.2.1 Results without Stopwords Removal

In the first experimental scenario, preprocessing was conducted without stopwords removal to examine the influence of stopwords on model performance in detecting clickbait in Indonesian news headlines. This approach aimed to retain all words in the dataset, including common and less informative words, to observe their impact on model performance. The evaluation results of the baseline models, consisting of Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support Vector Machine (SVM), are presented in Table 4. The table provides the accuracy, precision, recall, and F1-score metrics for each fold during the 5-fold cross-validation process.

Based on Table 4, the Multinomial Naive Bayes model achieved an average accuracy of 0.7543, a precision of 0.7450, and a recall of 0.6297. The F1-score of 0.6825 indicates a balanced trade-off between precision and recall. However, the model's performance varied across the folds, with fold 4 recording the lowest recall and F1-score, suggesting a tendency for the model to misclassify clickbait as non-clickbait. The Logistic Regression model demonstrated improved performance, with an average accuracy of 0.7717 and the highest average precision of 0.7884. Nevertheless, the recall remained relatively low at 0.6226, indicating that although the model effectively minimized false positives, it still struggled to comprehensively capture clickbait patterns. The SVM model exhibited the best performance among the baseline models, achieving an average accuracy of 0.7734 and an F1-score of 0.7016. The SVM maintained precision stability across all folds, with the highest precision of 0.7888 observed in fold 3. In contrast, the model's recall fluctuated, particularly in fold 4, where it dropped to 0.6343. This decline in recall suggests that despite SVM's high overall performance, it still encountered challenges in consistently identifying clickbait instances.

Table 4 Performance Metrics of Baseline Model without Stopwords Removal

Model	Fold	Accuracy	Precision	Recall	F1-Score
MultinomialNB	1	0.7533	0.7467	0.6232	0.6794
	2	0.7590	0.7443	0.6479	0.6927
	3	0.7693	0.7690	0.6431	0.7004
	4	0.7403	0.7240	0.6153	0.6652
	5	0.7497	0.7412	0.6192	0.6748
	Avg	0.7543	0.7450	0.6297	0.6825
Logistic Regression	1	0.7730	0.7899	0.6248	0.6977
	2	0.7767	0.7946	0.6304	0.7030
	3	0.7743	0.7902	0.6288	0.7003
	4	0.7683	0.7852	0.6161	0.6904
	5	0.7660	0.7819	0.6129	0.6872
	Avg	0.7717	**0.7884**	0.6226	0.6957
SVM	1	0.7750	0.7861	0.6367	0.7036
	2	0.7777	0.7911	0.6383	0.7066
	3	0.7793	0.7888	0.6471	0.7109
	4	0.7713	0.7793	0.6343	0.6994
	5	0.7637	0.7710	0.6208	0.6878
	Avg	**0.7734**	0.7832	**0.6355**	**0.7016**

Table 5 Performance Metrics of Ensemble Models without Stopwords Removal.

Model	Fold	Accuracy	Precision	Recall	F1-Score
Voting Classifier	1	0.7743	0.7818	0.6407	0.7042
	2	0.7800	0.7842	0.6558	0.7143
	3	0.7853	0.7929	0.6606	0.7207
	4	0.7730	0.7771	0.6431	0.7038
	5	0.7690	0.7745	0.6335	0.6970
	Avg	**0.7763**	0.7821	**0.6467**	**0.7080**
Stacking Classifier	1	0.7760	0.7878	0.6375	0.7047
	2	0.7773	0.7921	0.6359	0.7055
	3	0.7770	0.7896	0.6383	0.7059
	4	0.7703	0.7831	0.6256	0.6955
	5	0.7633	0.7708	0.6200	0.6872
	Avg	0.7728	**0.7847**	0.6315	0.6998

The use of ensemble learning methods enhanced model performance, as demonstrated in Table 5. The Voting Classifier achieved an average accuracy of 0.7763, outperforming all baseline models. The F1-score of 0.7080 indicates that the integration of predictions from base learners effectively balanced precision and recall, reducing the impact of false positives and false negatives. However, similar to the baseline models, the recall remained lower than the precision, indicating the model's difficulty in accurately capturing clickbait instances. The ensemble model based on the Stacking Classifier exhibited lower accuracy compared to the Voting Classifier. Despite achieving a high precision of 0.7847, the relatively low recall of 0.6315 suggests that the meta-learner may have overly relied on the base models with high precision, potentially neglecting patterns indicative of clickbait that could have been identified with a more balanced approach.

4.2.2 Results with Stopwords Removal

In the second scenario, preprocessing was conducted by applying stopwords removal to reduce noise in the data and retain words deemed more informative. This approach aimed to eliminate common words such as conjunctions, prepositions, and adverbs, which are generally considered to have minimal contributions to detecting clickbait patterns in Indonesian news headlines. The evaluation results of the baseline models, consisting of Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support Vector Machine (SVM), are presented in Table 6. The table provides accuracy, precision, recall, and F1-score metrics for each fold during the 5-fold cross-validation process.

Table 6 Performance Metrics of Baseline Models with Stopwords Removal.

Model	Fold	Accuracy	Precision	Recall	F1-Score
MultinomialNB	1	0.7293	0.7226	0.5755	0.6407
	2	0.7310	0.7183	0.5898	0.6478
	3	0.7380	0.7261	0.6025	0.6586
	4	0.7233	0.7094	0.5763	0.6360
	5	0.7263	0.7123	0.5827	0.6410
	Avg	0.7296	0.7177	0.5854	0.6448
Logistic Regression	1	0.7343	0.7460	0.5556	0.6369
	2	0.7373	0.7433	0.5707	0.6457
	3	0.7443	0.7482	0.5882	0.6587
	4	0.7300	0.7333	0.5596	0.6348
	5	0.7330	0.7383	0.5628	0.6387
	Avg	0.7358	0.7418	0.5674	0.6429
SVM	1	0.7333	0.7267	0.5835	0.6473
	2	0.7310	0.7262	0.5755	0.6421
	3	0.7447	0.7356	0.6105	0.6672
	4	0.7267	0.7194	0.5707	0.6365
	5	0.7263	0.7128	0.5819	0.6407
	Avg	0.7324	0.7241	0.5844	0.6468

Based on Table 6, the Multinomial Naive Bayes model recorded a decline in average accuracy to 0.7296, with a precision of 0.7177 and a recall of 0.5854. The F1-score of 0.6448 indicates that the removal of stopwords adversely impacted recall, resulting in an increase in false negatives. Fold 4 exhibited the lowest recall of 0.5763, suggesting that without common words, the model became less effective in accurately detecting clickbait. Logistic Regression model also experienced a decline in performance after stopwords removal was applied. The average accuracy dropped to 0.7358, with a precision of 0.7418 and a recall of 0.5674. Although precision remained stable, the decrease in recall

indicates that the model became more susceptible to false negatives, particularly in fold 1, where recall dropped to 0.5556, the lowest among all folds. This suggests that removing common words may obscure clickbait patterns that were previously identifiable by the model. The SVM model displayed a similar trend, with an average accuracy of 0.7324 and an F1-score of 0.6468. Precision remained stable at 0.7241, but recall decreased to 0.5844, indicating a reduced ability of the model to detect clickbait without common words as contextual cues. Fold 4 again recorded the lowest recall at 0.5707, suggesting that while SVM maintained relatively stable precision, it struggled to maintain sensitivity to clickbait instances when stopwords were removed.

Table 7 Performance Metrics of Ensemble Models with Stopwords Removal.

Model	Fold	Accuracy	Precision	Recall	F1-Score
Voting Classifier	1	0.7370	0.7395	0.5755	0.6473
	2	0.7390	0.7354	0.5898	0.6546
	3	0.7420	0.7359	0.6002	0.6611
	4	0.7313	0.7274	0.5747	0.6421
	5	0.7363	0.7305	0.5882	0.6517
	Avg	0.7371	0.7337	0.5857	0.6514
Stacking Classifier	1	0.7327	0.7562	0.5350	0.6266
	2	0.7430	0.7621	0.5628	0.6475
	3	0.7447	0.7611	0.5700	0.6518
	4	0.7293	0.7511	0.5302	0.6216
	5	0.7333	0.7468	0.5509	0.6340
	Avg	0.7366	0.7555	0.5498	0.6363

The application of ensemble learning methods also demonstrated a decline in performance after stopwords removal, as shown in Table 7. The Voting Classifier recorded an accuracy of 0.7371, slightly lower than the scenario without stopwords removal. The F1-score of 0.6514 indicates that although precision remained high at 0.7337, the reduction in recall from 0.6467 to 0.5857 led to an increase in false negatives. The Stacking Classifier experienced a more decline in performance, with accuracy dropping to 0.7366 and F1-score to 0.6363. Precision reached 0.7555, but the low recall of 0.5498 suggests that the meta-learner Logistic Regression relied too heavily on base models with high precision, while its sensitivity to clickbait patterns decreased due to the removal of stopwords.

4.2.3 Impact of Stopwords Removal

The impact of stopwords removal on model performance is illustrated in Figure 4. The figure shows a noticeable decline in average accuracy across all models after stopwords removal.

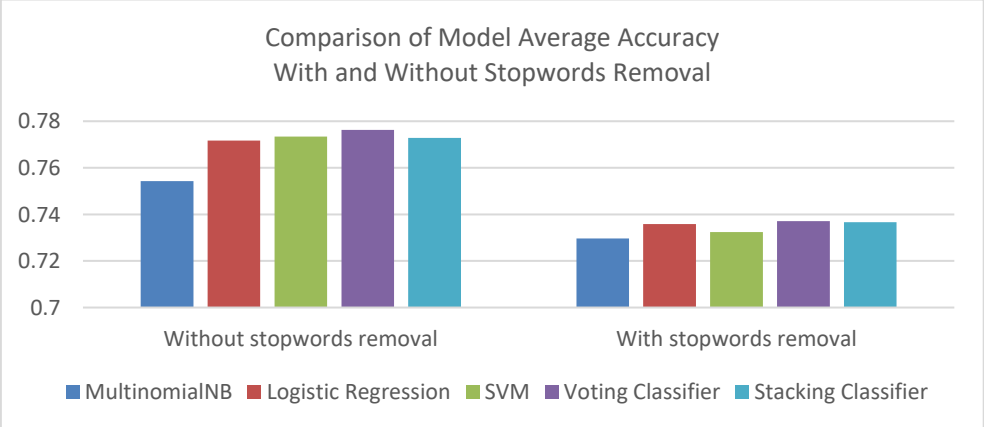

Figure 4 Comparison of Model Accuracy With and Without Stopwords Removal

For the baseline models, Multinomial Naive Bayes exhibited the most significant reduction in accuracy, decreasing from 0.7543 to 0.7296. Logistic Regression and SVM also experienced decreases in accuracy, from 0.7717 to 0.7358 and from 0.7734 to 0.7324, respectively. The decline suggests that the removal of common words reduced the models' ability to effectively identify clickbait patterns, potentially obscuring contextual cues that are essential in distinguishing clickbait from non-clickbait content. Similarly, the ensemble methods also showed lower accuracy with stopwords removal. The Voting Classifier and Stacking Classifier, which previously achieved the highest accuracies without stopwords removal (0.7763 and 0.7728, respectively), recorded lower accuracies of 0.7371 and 0.7366 after stopwords were removed. This decline indicates that the ensemble strategies, despite integrating predictions from multiple base learners, struggled to maintain their robustness without the contextual information provided by common words.

Overall, the decrease in accuracy across all models highlights the potential loss of valuable contextual information due to stopwords removal, particularly in clickbait detection, where common words may serve as significant indicators of sensationalist or attention-grabbing content.

4.2.4 Impact of Ensemble Learning

The impact of ensemble learning on model accuracy is illustrated in Figure 5. The figure clearly shows that the ensemble models, Voting Classifier and Stacking Classifier, achieved higher accuracy compared to the baseline models, demonstrating the effectiveness of integrating predictions from multiple base learners.

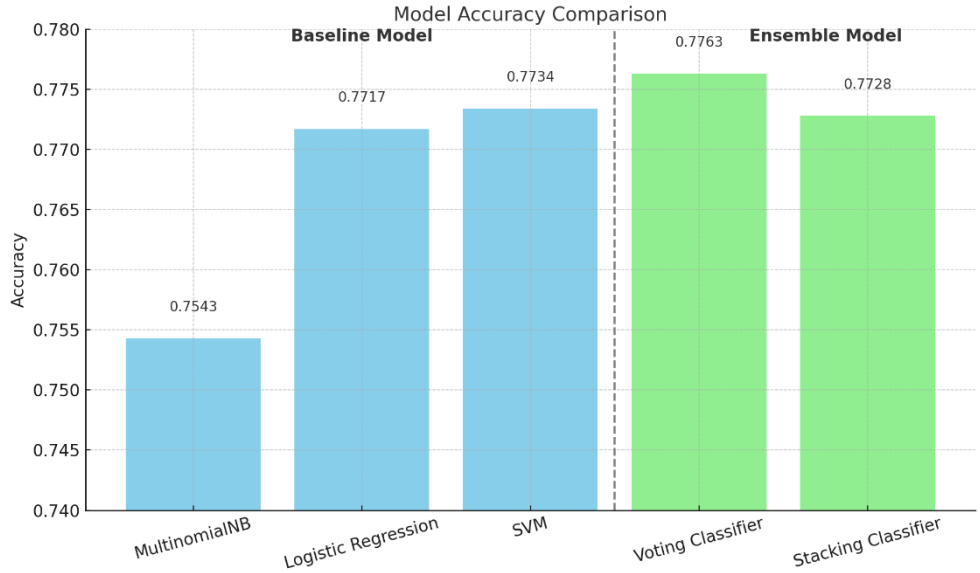

Figure 5 Model Accuracy Comparison Between Baseline and Ensemble Models

The Voting Classifier achieved the highest accuracy of 0.7763, outperforming all baseline models, including SVM, which recorded the highest accuracy among the baseline models at 0.7734. This improvement suggests that the soft voting mechanism effectively leveraged the strengths of each base learner, reducing the impact of individual model weaknesses and enhancing overall performance. The Stacking Classifier also exhibited a notable increase in accuracy, reaching 0.7728, slightly lower than the Voting Classifier but still higher than the baseline models. The use of Logistic Regression as

the meta-learner allowed the stacking model to combine the predictions of the base learners more comprehensively, albeit with slightly lower accuracy compared to the Voting Classifier.

Table 8 Evaluation of classification model performance based on precision, recall, and F1-score per class

Model	Class	Precision	Recall	F1-Score
MultinomialNB	Non-clickbait	0.7595	0.8443	0.7996
	Clickbait	0.7450	0.6297	0.6825
Logistic Regression	Non-clickbait	0.7634	0.8793	0.8173
	Clickbait	0.7884	0.6226	0.6957
SVM	Non-clickbait	0.7683	0.8730	0.8173
	Clickbait	0.7832	0.6355	0.7016
Voting Classifier	Non-clickbait	0.7733	0.8699	0.8187
	Clickbait	0.7821	0.6467	0.7080
Stacking Classifier	Non-clickbait	0.7668	0.8749	0.8172
	Clickbait	0.7847	0.6315	0.6998

Table 8 presents the evaluation results of five classification models based on precision, recall, and F1-score for each class. Overall, all models achieve higher performance in identifying Non-clickbait headlines compared to Clickbait. This is evident from the consistently higher recall and precision values for the Non-clickbait class. Among the models, the Voting Classifier achieved the best overall performance, especially in balancing precision (0.7821) and recall (0.6467) for the Clickbait class, resulting in the highest F1-score (0.7080) for that category.

The precision values for the Clickbait class across all models are slightly lower, indicating a moderate rate of false positives, headlines incorrectly labeled as clickbait. Meanwhile, the recall values for the Clickbait class are notably lower than for Non-clickbait, suggesting that many true clickbait headlines remain undetected. This trend implies that the models are more conservative in labeling a headline as clickbait, possibly due to overlapping linguistic features between the two classes. In contrast, Non-clickbait recall values are very high, with Logistic Regression, SVM, and Stacking achieving recall scores above 0.87. This means that the models are highly effective at correctly identifying Non-clickbait headlines, but this comes at the cost of reduced sensitivity to detecting Clickbait. The trade-off observed here reflects the models' tendency to favor precision over recall for the clickbait class, which may be appropriate depending on the application context.

Overall, the results indicate that the use of ensemble learning effectively improved model performance, particularly in maintaining accuracy. However, the relatively small gap between the ensemble models and the best-performing baseline model (SVM) suggests that further optimization of the meta-learner and the selection of base models may be necessary to maximize the benefits of ensemble learning in clickbait detection.

4.2.5 Model and Missclassification

In this study, we consider false negatives where clickbait headlines are misclassified as non-clickbait to be more critical. This type of error allows manipulative or sensational content to bypass detection, potentially harming user experience and reducing the credibility of content platforms. On the other hand, false positives may lead to the exclusion of informative headlines, which is undesirable but less damaging. This consideration is essential when evaluating model performance in real-world deployment, where the cost of undetected clickbait is often higher than that of false alarms.

The selection of the three base models, Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support Vector Machine (SVM), was based on the distinct classification principles they employ: probabilistic, linear, and margin-based, respectively. These differences contribute to prediction diversity, which is a critical component in ensemble learning. In the soft voting approach, the final output is determined by averaging the class probabilities from the three models. Meanwhile, stacking allows a meta-learner (Logistic Regression) to learn from a combination of the base models' predictions and the original features, enabled through the passthrough option. This enables the stacking model to capture more complex patterns that individual models may not be able to handle.

Based on the evaluation results, the ensemble models, particularly stacking, demonstrated a significant performance improvement. Nevertheless, there remain certain headlines that are difficult to classify accurately. Examples of misclassified cases are presented in Table 9.

Table 9 Missclassified Headlines Examples

No.	Headline	Predicted Label	Actual Label
1	Polisi Belanda Tembak Mati Istri dan 2 Putrinya Lalu Bunuh Diri (Dutch Police Officer Shoots Wife and 2 Daughters Dead, Then Commits Suicide)	Clickbait	Non-clickbait
2	Remaja Asal Asahan Alami Kebutaan, Ngaku Keseringan Main Game Online (Teenager from Asahan Goes Blind, Claims to Play Online Games Excessively)	Clickbait	Non-clickbait
3	Warga Tanjungbalai Sumut Heboh Temukan 4 Buaya di Sungai (Residents of Tanjungbalai North Sumatra Shocked to Find 4 Crocodiles in River)	Clickbait	Non-clickbait
4	Mengenaskan! 15 Orang Sekeluarga Tewas dalam Serangan Arab Saudi di Yaman (Tragic! 15 Family Members Killed in Saudi Attack on Yemen)	Non-clickbait	Clickbait
5	Anggota DPRD Banten Gadai SK, Komisi II DPR: Tutup Utang Kampanye? (Banten Regional Parliament Member Pawns Decree, House Commission II: Covering Campaign Debt?)	Non-clickbait	Clickbait

The analysis of five misclassified cases reveals the challenges faced by the model in recognizing context and emotional intensity. Headlines 1 to 3 were predicted as clickbait despite being labeled as non-clickbait. This may be attributed to the presence of sensational or shocking words—such as those referring to tragic or dramatic events—which may lead the model to assume clickbait intent. Conversely, headlines 4 and 5, which were in fact clickbait, were classified as non-clickbait. This indicates that the model may be less responsive to explicit clickbait characteristics, such as the use of exclamation marks or speculative phrasing. These findings highlight that while the ensemble model performs well overall, further refinement is needed to better capture subtle nuances in language.

5 Conclusion

The experimental results in this study indicate that retaining stopwords generally led to higher accuracy and F1-scores across all models. This suggests that stopwords may carry significant contextual value in identifying clickbait patterns in news headlines. In contrast, the removal of stopwords resulted in a decline in accuracy for all models, with SVM experiencing the most substantial drop of 0.0410. Other models also recorded notable decreases of 0.0392, 0.0362, 0.0359, and 0.0247 for the Voting Classifier, Stacking Classifier, Logistic Regression, and Multinomial Naive Bayes, respectively. Despite the decline, the Voting Classifier achieved the highest accuracy among all models, indicating that the use of ensemble learning can enhance model performance compared to its baseline counterparts. However, the relatively small performance gap between ensemble models and baseline models suggests that further performance gains could potentially be achieved by optimizing the selection of meta-learners and base learners in the ensemble model.

Bibliography

[1] X. Li, J. Zhou, H. Xiang, and J. Cao, "Attention Grabbing through Forward Reference: An ERP Study on Clickbait and Top News Stories," *Int J Hum Comput Interact*, vol. 40, no. 11, pp. 3014–3029, 2024, doi: 10.1080/10447318.2022.2158262.

[2] A.-K. Jung, S. Stieglitz, T. Kissmer, M. Mirbabaie, and T. Kroll, "Click me. . .! The influence of clickbait on user engagement in social media and the role of digital nudging," *PLoS One*, vol. 17, no. 6 June, 2022, doi: 10.1371/journal.pone.0266743.

[3] L. S. Shiang and S. Wilson, "Unravelling clickbait news as viral journalism in Malaysia: Its phenomenon and impacts," *SEARCH Journal of Media and Communication Research*, vol. 16, no. 1, pp. 33–47, 2024.

[4] K. Janét, O. Richards, and A. R. Landrum, "Headline Format Influences Evaluation of, but Not Engagement with, Environmental News," *Journalism Practice*, vol. 16, no. 1, pp. 35–55, 2022, doi: 10.1080/17512786.2020.1805794.

[5] A. William and Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," *Data Brief*, vol. 32, p. 106231, Oct. 2020, doi: 10.1016/J.DIB.2020.106231.

[6] S. Kurniawan and I. Budi, "Indonesian Tweets Hate Speech Target Classification Using Machine Learning," in *2020 5th International Conference on Informatics and Computing, ICIC 2020*, 2020, pp. 1–5. doi: 10.1109/ICIC50835.2020.9288515.

[7] P. Klairith and S. Tanachutiwat, "Thai clickbait detection algorithms using natural language processing with machine learning techniques," in *ICEAST 2018 - 4th International Conference on Engineering, Applied Sciences and Technology: Exploring Innovative Solutions for Smart Society*, IEEE, 2018, pp. 1–4. doi: 10.1109/ICEAST.2018.8434447.

[8] C. Wu, F. Wu, T. Qi, and Y. Huang, *Clickbait Detection with Style-Aware Title Modeling and Co-attention*, vol. 12522 LNAI. 2020. doi: 10.1007/978-3-030-63031-7_31.

[9] M. Zhou, W. Xu, W. Zhang, and Q. Jiang, "Leverage knowledge graph and GCN for fine-grained-level clickbait detection," *World Wide Web*, vol. 25, no. 3, pp. 1243–1258, May 2022, doi: 10.1007/s11280-022-01032-3.

[10] K. K. Yadav and N. Bansal, "A Comparative Study on Clickbait Detection using Machine Learning Based Methods," in *2023 International Conference on Disruptive Technologies, ICDT 2023*, 2023, pp. 661–665. doi: 10.1109/ICDT57929.2023.10150475.

[11] D. S. Sisodia, "Ensemble learning approach for clickbait detection using article headline features," *Inf Sci*, vol. 22, no. 2019, pp. 31–44, 2019, doi: 10.28945/4279.

[12] R. Rajesh Sharma, A. Sungheetha, M. A. Haile, A. H. Kedir, A. Rajasekaran, and G. Charles Babu, "Clickbait Detection for Amharic Language Using Deep Learning Techniques," *Journal of Machine and Computing*, vol. 4, no. 3, pp. 603–615, 2024, doi: 10.53759/7669/jmc202404058.

[13] A. Chowanda, Nadia, and L. M. M. Kolbe, "Identifying clickbait in online news using deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1755–1761, 2023, doi: 10.11591/eei.v12i3.4444.

[14] B. U. Nadia and I. A. Iswanto, "Indonesian Clickbait Detection Using Improved Backpropagation Neural Network," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2021*, 2021, pp. 252–256. doi: 10.1109/ISRITI54043.2021.9702872.

[15] P. Santoso Hadi, Muljono, A. Z. Fanani, G. F. Shidik, Purwanto, and F. Alzami, "Using Extra Weight in Machine Learning Algorithms for Clickbait Detection of Indonesia Online News Headlines," in *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021*, 2021, pp. 37–41. doi: 10.1109/iSemantic52711.2021.9573213.

[16] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[17] S. Gan, S. Shao, L. Chen, L. Yu, and L. Jiang, "Adapting hidden naive bayes for text classification," *Mathematics*, vol. 9, no. 19, 2021, doi: 10.3390/math9192378.

[18] T.-T. Wong and H.-C. Tsai, "Multinomial naïve Bayesian classifier with generalized Dirichlet priors for high-dimensional imbalanced data," *Knowl Based Syst*, vol. 228, 2021, doi: 10.1016/j.knosys.2021.107288.

[19] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," in *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*, 2019, pp. 135–139. doi: 10.1109/ICCSNT47585.2019.8962457.

[20] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artif Intell Rev*, vol. 52, no. 2, pp. 803–855, 2019, doi: 10.1007/s10462-018-9614-6.

[21] T. Wang, F. Liu, and S. Yan, "Learning class-informed exponential kernel for text categorization," *J Comput Theor Nanosci*, vol. 13, no. 8, pp. 5103–5110, 2016, doi: 10.1166/jctn.2016.5389.

[22] M. K. Anam, M. B. Firdaus, F. Suandi, Lathifah, T. Nasution, and S. Fadly, "Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining," in *ICFTSS 2024 - International Conference on Future Technologies for Smart Society*, 2024, pp. 90–95. doi: 10.1109/ICFTSS61109.2024.10691363.

[23] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, *Voting and Bagging*, vol. 334. 2022. doi: 10.1007/978-3-031-16990-8_14.

[24] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren, "StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," *IEEE*

Trans Vis Comput Graph, vol. 27, no. 2, pp. 1547–1557, 2021, doi: 10.1109/TVCG.2020.3030352.

[25] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare (Switzerland)*, vol. 11, no. 12, 2023, doi: 10.3390/healthcare11121808.