International Journal of Scientific Engineering and Science
Volume 9, Issue 5, pp. 163-168, 2025. ISSN (Online): 2456-7361

Evaluation of Logistic Regression and Random Forest Algorithms for Hate Speech Identification

Adam Julizar[1], Mia Kamayani Sulaeman[2]
[1]Information Engineering, Prof. Dr. Hamka Muhammadiyah University, East Jakarta, Indonesia, 13830
[2] Information Engineering, Prof. Dr. Hamka Muhammadiyah University, East Jakarta, Indonesia, 13830
[1]2103015227@uhamka.ac.id, [2]mia.kamayani@uhamka.ac.id

***Abstract*—** *Hate speech on social media presents a substantial threat to digital well-being, particularly in linguistically diverse contexts such as Indonesia. This study aims to evaluate and compare the performance of two machine learning algorithms—Logistic Regression and Random Forest—for classifying Indonesian-language hate speech on social media platforms. A dataset of 11,122 annotated text entries was obtained from Kaggle and subjected to preprocessing steps, including text cleaning, normalization, stopword removal, and stemming. Two feature extraction approaches were explored: Term Frequency–Inverse Document Frequency (TF-IDF) alone and TF-IDF combined with N-Gram. The evaluation results show that combining TF-IDF with N-Gram improved the accuracy of Logistic Regression from 81% to 83% and Random Forest from 82% to 83%, with a notable improvement in Logistic Regression's recall from 79% to 85%. Random Forest exhibited more stable performance across scenarios, while Logistic Regression offered faster computation and easier interpretability. These findings suggest that TF-IDF + N-Gram is an effective feature extraction method for Indonesian hate speech detection, with Logistic Regression suitable for real-time systems and Random Forest preferred for accuracy-focused applications. The study contributes to the advancement of multilingual hate speech detection systems tailored for the Indonesian language.*

***Keywords*—** *Hate speech, Logistic Regression, Random Forest, TF-IDF, N-Gram.*

I. INTRODUCTION

1. Background

Over the past decade, social media usage has surged dramatically, with various major platforms—including leading networks such as Meta's flagship site, microblogging services, photo-sharing apps, and short-form video platforms—collectively engaging billions of active users globally. As of 2023, the worldwide number of active social media participants neared five billion, with people spending an average of 2 hours and 27 minutes per day on these platforms[1].

The rapid evolution of social media technology enables users to instantaneously share diverse content formats—including text, audio, images, and video—across geographical and temporal boundaries via internet connectivity. However, this ease of communication also presents opportunities for misuse, with some individuals exploiting these platforms to express negative sentiments, disseminate misinformation, discredit others, and propagate hate speech targeting individuals or specific groups[2].

Hate speech encompasses acts involving insult, defamation, blasphemy, provocation, incitement, the spread of false information, or other offensive behaviors that can incite discrimination, violence, or social conflict, and even endanger the lives of individuals or groups. Such actions, perpetrated by individuals or collectives depending on their intent, often involve slander and the manipulation of negative public opinion against victims, causing distress[3].

Hate speech on social media frequently violates norms of civil discourse and communication ethics. In this context, the phenomenon not only affects targeted individuals but can also escalate into broader societal conflicts. A particularly concerning aspect of online hate speech is its documented association with violent acts against members of targeted groups, as exemplified by the "Unite the Right" attacks in Charlottesville, the Pittsburgh synagogue shooting, and the Rohingya genocide in Myanmar, among other instances. Consequently, national governments and supranational organizations, such as the European Union, have enacted legislation urging social media companies to moderate and remove discriminatory content, with a specific focus on material inciting physical violence[4].

The impact of hate speech is substantial. Research indicates that individuals victimized by hate speech are more susceptible to mental health challenges, potentially experiencing psychological disorders such as anxiety, emotional distress, and fear of online threats materializing in real-world scenarios. Furthermore, continuous exposure to such behavior can lead to desensitization, normalizing hate speech within social interactions[5].

These findings highlight the critical importance of accurately detecting and classifying hate speech to mitigate its harmful effects. Numerous classification methods have been developed for this purpose, particularly on social media platforms. For instance, a study conducted by Khan proposed an automated hate speech detection technique for English-language content using six different combinations of machine learning and natural language processing (NLP). The study reported that the Logistic Regression algorithm, combined with TF-IDF feature extraction and n-gram techniques, achieved a classification accuracy of 94% in distinguishing hate speech from non-hate speech content[6].

However, research employing similar methodologies within the Indonesian language context remains limited. While some prior studies have applied machine learning algorithms to detect hate speech in Indonesian, there is a notable lack of

research explicitly comparing the performance of Logistic Regression and Random Forest algorithms under two feature extraction scenarios: TF-IDF alone and TF-IDF in combination with N-gram features. For instance, Riadi's study, which utilized Support Vector Machine (SVM) with TF-IDF for Indonesian hate speech detection, reported an accuracy of 84%. Nevertheless, this study did not include a comparison with Logistic Regression or investigate the integration of TF-IDF and N-gram techniques[7]. This highlights a significant research gap that necessitates further investigation to identify more accurate and efficient classification approaches, particularly within the domain of natural language processing for the Indonesian language.

Given the increasing awareness of the dangers posed by hate speech and the advancements in data analysis technologies, this study aims to further explore effective classification methods for detecting hate speech in Indonesian on social media platforms. This research is expected to contribute to the development of more comprehensive and sustainable solutions for addressing this critical issue.

II. Literature Review

2.1 Hate Speech:

A complex and multidimensional phenomena, hate speech (HS) has attracted scholarly study from a wide range of fields, including psychology, sociology, communication studies, and law. According to its broad definition, it is any deliberate and deliberate public comment meant to disparage, degrade, or provoke disdain against people or groups on the basis of distinguishing traits including color, ethnicity, religion, gender, sexual orientation, or nationality. Hate speech, as further explained in the European Commission's Recommendation against Racism and Intolerance, includes statements of hatred, humiliation, or contempt aimed against an individual or group, frequently upholding social hierarchies and structural injustices[8].

2.2 Logistic Regression:

The logistic function serves as the primary method for representing a binary dependent variable in statistical analyses known as logistic regression. Logistic Regression (LR) aims to establish a relationship between a set of independent variables and the outcome by applying an appropriate model. The confidence level of an outcome in the LR model is calculated using the logistic function Q(x), which depends on the independent variables. This function, also referred to as the sigmoid function, generates outputs in the range [0,1] based on real-valued inputs. These outputs are typically interpreted as probabilities. The result indicates the model's confidence in the classification, with values near 0 representing the negative class and values close to 1 representing the positive class. The expression is presented in Equation[9]:

$$Q(x) = \frac{1}{1 + e^{-x}}$$

2.3 Random Forest:

In numerous research contexts, predictive models are created using the random forest machine learning technique. The text corpus is a well-known example of the kind of high-dimensional data that our technique is specifically made to investigate. As a result, one method that can be used to analyze sentiment is random forest. Using a randomly chosen subset of samples and training factors, the Random Forest ensemble classifier produces a large number of decision tree algorithms. Random procedures are used to ensure that the decision tree algorithms don't influence one another[10].

2.4 TF-IDF:

A popular feature weighting method that uses numerical statistical techniques to determine and measure the significance of terms within a collection of documents is the Term Frequency–Inverse Document Frequency (TF-IDF) method. The core idea behind TF-IDF is the calculation of two essential elements: Inverse Document Frequency (IDF), which assesses a term's originality or rarity throughout the entire corpus, and Term Frequency (TF), which measures a term's frequency within a particular document. This combination enables TF-IDF to identify phrases that are both unique throughout the dataset and frequently occurring in a particular document. Equation provides the TF-IDF mathematical formulation[11]:

$$W_{t,d} = tf_{t,d} * \log\left(\frac{N}{df_t}\right)$$

Wich $W_{t,d}$ is the weight of term in document d, $tf_{t,d}$ is value of TF and df_t is the number of document containing word t

2.5 N-Gram:

N-Gram is a technique that considers the order of words in a text by forming a combination of consecutive words of n words. For example, the bigram (n=2) of the sentence "I like reading" is ["I like", "like reading"]. The use of N-Gram allows the model to capture local context and relationships between adjacent words, which are often important in understanding the meaning of the text as a whole (Priyatno & Firmananda, 2022). This study uses N-gram words with specifications, namely bigrams[12].

III. Methodology

This section will explain the design of the hate speech detection system. This design consists of several steps, data collection, data preprocessing, Feature Extraction (TF-IDF, N-gram), data classification (LR, RF) and system evaluation using the confusion matrix method. Figure 1 will explain the system flow.

3.1 Data Collection

This study utilizes a dataset obtained from the Kaggle platform, developed by Ibrohim and Budi. The dataset comprises 11,122 entries of annotated Indonesian-language text, which have been categorized into two labels: hate speech (label 1) and non-hate speech (label 0). The selection of this dataset was based on its open accessibility, the relevance of its content to the research topic, and the representativeness of the data in reflecting the phenomenon of hate speech on social media[13].

International Journal of Scientific Engineering and Science
Volume 9, Issue 5, pp. 163-168, 2025. ISSN (Online): 2456-7361

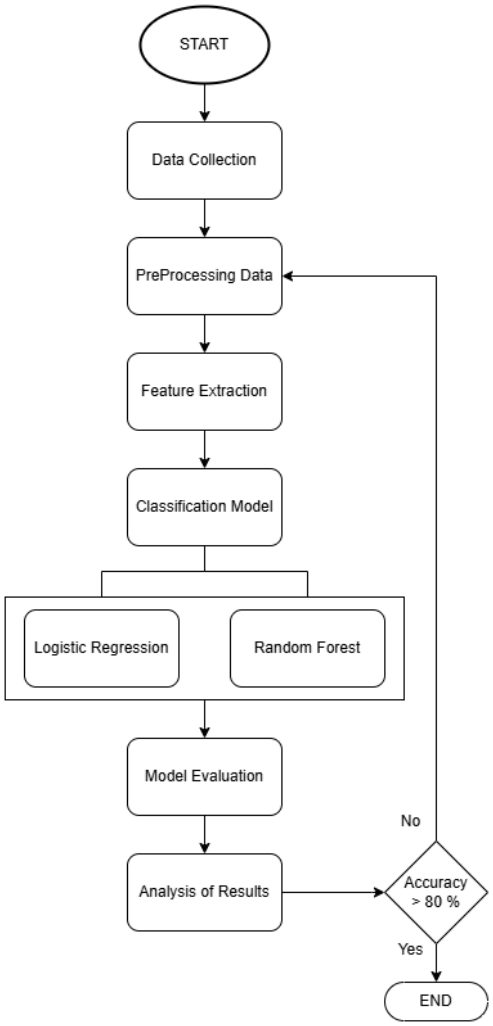

Fig. 1. system flow

3.2 Data Preprocessing

Data preprocessing aims to enhance the quality of textual data to ensure it is well-prepared and suitable for training classification models. The preprocessing steps were conducted systematically as follows:

- Text Cleaning: Irrelevant elements such as URLs, meaningless numbers, punctuation marks, special characters (e.g., @, #, $, %, etc.), and excessive whitespace were removed[14].
- Lowercasing: All text was converted to lowercase to standardize the format and prevent feature duplication due to capitalization[15].
- Normalization of Informal Words: Informal or non-standard words (commonly known as gaul terms) were converted into their standard forms using a specialized Indonesian normalization dictionary[16].
- Stopword Removal: Common words that do not significantly contribute to semantic analysis, such as and, which, and at, were eliminated[17].
- Stemming: Affixed words were reduced to their root forms using the Sastrawi stemming algorithm, ensuring consistent representation of each term in its base form[18].

All preprocessing steps were performed using the Python programming language with the support of natural language processing libraries, including Sastrawi, re, and a custom Indonesian normalization dictionary.

3.3. Feature Extraction

After text preprocessing, the textual data were transformed into numerical representations using two approaches:

- TF-IDF (Term Frequency–Inverse Document Frequency): This method measures the importance of a word in a document relative to the entire corpus[19].
- TF-IDF Combined with N-Gram: In addition to single-word tokens (unigrams), this approach also incorporates two-word combinations (bigrams) to capture more complex contextual information[20].

Both methods were employed to evaluate the impact of feature extraction techniques on classification performance.

3.4. Model Classification

This study implements two machine learning algorithms:

- Logistic Regression: A regression-based model used to predict binary outcomes, suitable for text classification tasks[21].
- Random Forest: A tree-based ensemble algorithm capable of efficiently handling large datasets while minimizing the risk of overfitting by constructing multiple decision trees using random subsets of data and aggregating their outcomes[22].

Each model was trained under two scenarios: (1) using TF-IDF only and (2) using a combination of TF-IDF and N-Gram.

3.5 Evaluation

To evaluate model performance, several evaluation metrics were employed, including accuracy, precision, recall, and F1-score[23].

- Accuracy represents the proportion of correct predictions, considering both true positives and true negatives, derived from a comprehensive dataset.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} x\ 100\%$$

- F1 Score is an evaluation metric that combines precision and recall into a single balanced measure using their harmonic mean.

$$f1\ Score = \frac{2xPrecisionxRecall}{Precision + Recall}$$

- Recall This value represents the number of correctly predicted positive instances out of all the truly positive instances.

$$Recall = \frac{TP}{(FN+TP)} \text{ x } 100\%$$

- Precision This value represents the number of correctly predicted positive instances divided by the total number of instances predicted as positive.

$$Precision = \frac{TP}{(FP + TP)} X\ 100\%$$

Using a train-test split method, the training and evaluation procedure divided the data into 80% for training and 20% for testing.

International Journal of Scientific Engineering and Science
Volume 9, Issue 5, pp. 163-168, 2025. ISSN (Online): 2456-7361

3.6 Analysis of Result

After completing the evaluation of the classification models, the next step involves analyzing the results to understand the model's performance in detecting hate speech. The use of a confusion matrix for visualization, as well as the analysis of evaluation measures like accuracy, precision, recall, and F1-score, are all included in this analysis. If the evaluation results indicate that the model's accuracy is below 80%, the process will be repeated from the data preprocessing stage to make the necessary adjustments. However, if the model achieves an accuracy of 80% or higher, it is considered to have satisfactory performance and is deemed ready for deployment in hate speech detection applications.

IV. Result and Discussion

4.1 Model Performance Results

Two feature extraction techniques—TF-IDF alone and TF-IDF in conjunction with N-Gram—were used to assess the classification performance of two machine learning algorithms, Random Forest (RF) and Logistic Regression (LR). This study's evaluation metrics include F1-score, recall, accuracy, and precision.

TABLE 1. Performance Comparison of Logistic Regression and Random Forest

Algorithm	Feature Extraction	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	TF-IDF	81	81.5	81	81
Random Forest	TF-IDF	82	82.0	82	82
Logistic Regression	TF-IDF + N-Gram	83	83.0	83	83
Random Forest	TF-IDF + N-Gram	83	83.0	83	83

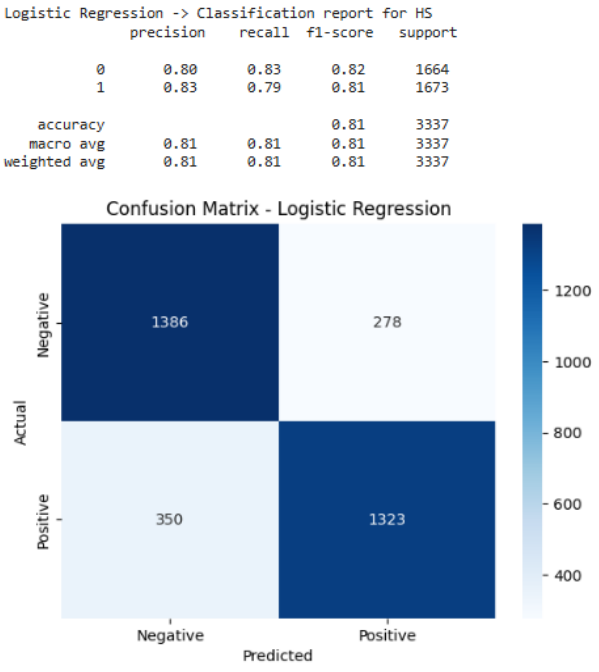

Fig. 2. Logistic Regression (TF-IDF)

The results indicate that the addition of N-Gram features leads to performance improvement in both algorithms. Notably, Logistic Regression exhibited a recall improvement from 79% to 85%, highlighting its increased ability to correctly identify hate speech instances after incorporating contextual word relationships.

4.2 Confusion Matrix Visualization

To better understand the classification performance, confusion matrices were generated for each model and feature combination. These are shown in Figures 2 through 5.

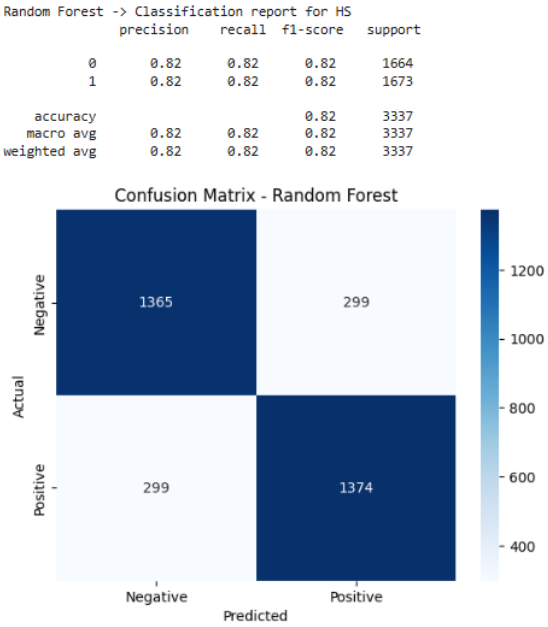

Fig. 3. Random Forest (TF-IDF)

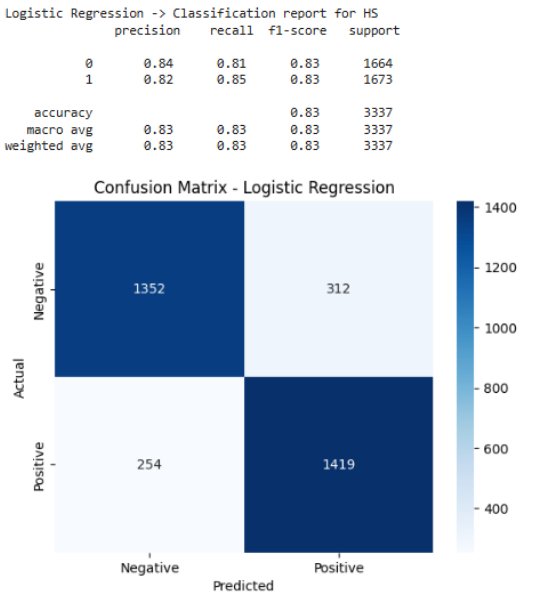

Fig. 4. Logistic Regression (TF-IDF+N-GRAM)

International Journal of Scientific Engineering and Science
Volume 9, Issue 5, pp. 163-168, 2025. ISSN (Online): 2456-7361

Random Forest -> Classification report for HS

	precision	recall	f1-score	support
0	0.84	0.80	0.82	1664
1	0.81	0.85	0.83	1673
accuracy			0.83	3337
macro avg	0.83	0.83	0.83	3337
weighted avg	0.83	0.83	0.83	3337

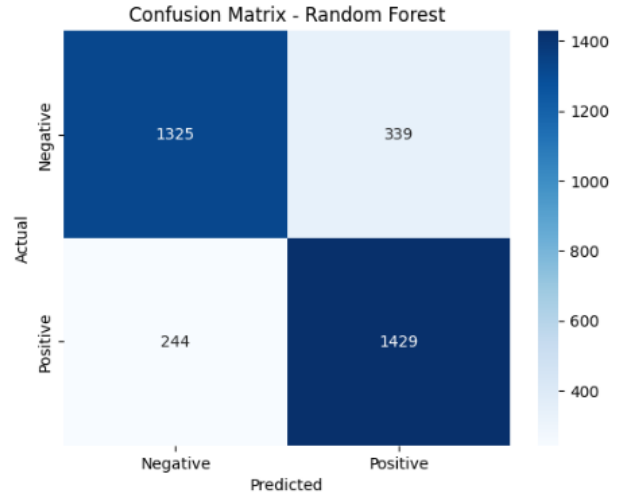

Fig. 5. Random Forest (TF-IDF+N-GRAM)

These confusion matrices provide a visual breakdown of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Improvements in TP and reductions in FN are particularly visible in the models utilizing the TF-IDF + N-Gram combination, demonstrating enhanced generalization in classification.

4.3 Comparative Analysis

The comparative results are summarized as follows:

- Accuracy: Both algorithms achieved the highest accuracy (83%) with the TF-IDF + N-Gram setting.
- Recall: Logistic Regression experienced a notable increase in recall from 79% to 85%, indicating its improved sensitivity in detecting hate speech.
- Stability: Random Forest demonstrated consistent performance across both feature settings, maintaining high precision and F1-score.
- Interpretability & Speed: Logistic Regression is preferred in scenarios requiring real-time processing and model interpretability due to its simpler linear nature.

TABLE 2. Summary of Model Strengths

Criteria	Logistic Regression	Random Forest
Best Accuracy	83% (TF-IDF + N-Gram)	83% (TF-IDF + N-Gram)
Best Recall	85% (TF-IDF + N-Gram)	85% (TF-IDF + N-Gram)
Interpretability	High	Moderate
Training Time	Fast	Slower
Stability	Slightly lower	More stable across features

These findings suggest that while both models benefit from contextual feature engineering, the choice between them depends on the application: Logistic Regression is ideal for real-time or resource-constrained systems, whereas Random Forest is better suited for tasks requiring more robust and consistent performance.

4.4 Discussion

The integration of N-Gram features notably enhanced the performance of both classification models by effectively capturing short-range contextual patterns frequently found in informal Indonesian language. In contrast, relying solely on TF-IDF, while computationally simpler, does not sufficiently capture linguistic depth, as indicated by slightly reduced recall metrics.

This study reinforces the importance of feature engineering in improving the effectiveness of machine learning models for text classification tasks—particularly in the context of detecting hate speech in morphologically complex languages such as Bahasa Indonesia. Furthermore, the balanced distribution of class labels (5561 hate speech and 5561 non-hate speech) enabled unbiased and consistent model evaluation.

Unlike many previous studies on Indonesian hate speech detection that typically apply a single algorithm or limited feature set, this research introduces a comparative analysis highlighting the real-world consequences of model and feature selection. It provides a valuable foundation for advancing scalable and adaptable hate speech detection systems, especially for under-resourced language settings.

V. CONCLUSION

This research investigated the effectiveness of two supervised machine learning algorithms—Logistic Regression and Random Forest—in identifying hate speech in Indonesian-language social media content. The study applied two feature extraction techniques: TF-IDF and a combination of TF-IDF with N-Gram. The dataset used included 11,122 labeled textual instances from Kaggle, and the preprocessing phase encompassed steps such as text cleaning, normalization, stopword elimination, and stemming.

The findings revealed that incorporating contextual features through the TF-IDF + N-Gram method led to improved accuracy for both classification models. Logistic Regression showed an increase in accuracy from 81% to 83%, while Random Forest also achieved an 83% accuracy. Additionally, the recall metric for Logistic Regression improved significantly from 79% to 85%, indicating a better ability to detect hate speech.

While both algorithms reached similar levels of accuracy, Random Forest offered greater performance consistency, whereas Logistic Regression provided faster execution and better interpretability. Based on these observations, Random Forest is recommended for tasks prioritizing classification precision, while Logistic Regression is more appropriate for systems requiring rapid response and transparency.

In summary, this study confirms the advantages of integrating contextual feature extraction with machine learning in enhancing hate speech detection in the Indonesian language. Future work may benefit from exploring deep learning models or multilingual datasets to expand classification performance further.

REFERENCES

[1] J. J. Van Bavel, C. E. Robertson, K. Del Rosario, J. Rasmussen, and S. Rathje, "Annual Review of Psychology Social Media and Morality," 2025, doi: 10.1146/annurev-psych-022123.

[2] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, no. 1, pp. 157–179, Jan. 2021, doi: 10.1007/s11192-020-03737-6.

[3] M. A. Paz, J. Montero-Díaz, and A. Moreno-Delgado, "Hate Speech: A Systematized Review," 2020, *SAGE Publications Inc.* doi: 10.1177/2158244020973022.

[4] J. M. Perez *et al.*, "Assessing the Impact of Contextual Information in Hate Speech Detection," *IEEE Access*, vol. 11, pp. 30575–30590, 2023, doi: 10.1109/ACCESS.2023.3258973.

[5] A. Stechemesser, A. Levermann, and L. Wenz, "Temperature impacts on hate speech online: evidence from 4 billion geolocated tweets from the USA," *Lancet Planet Health*, vol. 6, no. 9, pp. e714–e725, Sep. 2022, doi: 10.1016/S2542-5196(22)00173-5.

[6] J. Yousaf *et al.*, "HATE SPEECH DETECTION USING MACHINE LEARNING AND N-GRAM TECHNIQUES," May 2023. [Online]. Available: https://www.researchgate.net/publication/370492124

[7] I. Riadi, A. Fadlil, and U. Ahmad Dahlan Yogyakarta, "Identifying Hate Speech in Tweets with Sentiment Analysis on Indonesian Twitter Utilizing Support Vector Machine Algorithm," 2023.

[8] M. Hietanen and J. Eddebo, "Towards a Definition of Hate Speech—With a Focus on Online Contexts," *Journal of Communication Inquiry*, vol. 47, no. 4, pp. 440–458, Oct. 2023, doi: 10.1177/01968599221124309.

[9] A. A. Sosimi, O. Ipinnimo, C. O. Folorunso, B. A. Adim, and E. Onoyom-Ita, "HATE SPEECH IDENTIFICATION IN WEST AFRICA, USING MACHINE-LEARNING TECHNIQUES," vol. 20, no. 2, pp. 491–508, 2024, [Online]. Available: www.azojete.com.ng

[10] D. Indra Wijaya and R. Arifudin, "Detecting Hate Speech Tweets And Abusive Tweets In Indonesian Language Using Random Forest And Support Vector Machine With Voting Classifier Technique," *Journal of Advances in Information Systems and Technology*, vol. 4, no. 1, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jaist

[11] M. P. K. Dewi and E. B. Setiawan, "Feature Expansion Using Word2vec for Hate Speech Detection on Indonesian Twitter with Classification Using SVM and Random Forest," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, p. 979, Apr. 2022, doi: 10.30865/mib.v6i2.3855.

[12] A. M. Priyatno and F. I. Firmananda, "N-Gram Feature for Comparison of Machine Learning Methods on Sentiment in Financial News Headlines," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 1, no. 1, pp. 01–06, Jul. 2022, doi: 10.31004/riggs.v1i1.4.

[13] I. B. Muhammad Okky Ibrohim, "https://www.kaggle.com/code/wahyurizaldi/indonesian-twitter-hate-speech-text-analysis," https://www.kaggle.com/.

[14] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing Arabic text on social media," *Heliyon*, vol. 7, no. 2, Feb. 2021, doi: 10.1016/j.heliyon.2021.e06191.

[15] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and Suicide Analysis Using Machine Learning and NLP," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2022. doi: 10.1088/1742-6596/2161/1/012034.

[16] Indra, Agus Umar Hamdani, Suci Setiawati, Zena Dwi Mentari, and Mauridhy Hery Purnomo, "Comparison of K-NN, SVM, and Random Forest Algorithm for Detecting Hoax on Indonesian Election 2024," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 1, Mar. 2024, doi: 10.23887/janapati.v13i1.76079.

[17] S. Albahra *et al.*, "Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts," Mar. 01, 2023, *W.B. Saunders*. doi: 10.1053/j.semdp.2023.02.002.

[18] S. B. S, M. Y. M, D. Khyani, N. N. M, and D. B. M, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," Mar. 2021. [Online]. Available: https://www.researchgate.net/publication/348306833

[19] M. George and R. Murugesan, "Improving sentiment analysis of financial news headlines using hybrid Word2Vec-TFIDF feature extraction technique," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1–8. doi: 10.1016/j.procs.2024.10.172.

[20] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Sci J*, vol. 23, no. 1, 2024, doi: 10.5334/dsj-2024-031.

[21] G. Aliman *et al.*, "35 Sentiment Analysis using Logistic Regression Sentiment Analysis using Logistic Regression," 2022.

[22] M. A. Nivedha and S. Raja, "Detection of Email Spam using Natural Language Processing Based Random Forest Approach," *International Journal of Computer Science and Mobile Computing*, vol. 11, no. 2, pp. 7–22, Feb. 2022, doi: 10.47760/ijcsmc.2022.v11i02.002.

[23] N. P. Damayanti, D. E. Prameswari, W. Puspita, and P. S. Sundari, "Classification of Hate Comments on Twitter Using a Combination of Logistic Regression and Support Vector Machine Algorithm," *Journal of Information System Exploration and Research*, vol. 2, no. 1, Jan. 2024, doi: 10.52465/joiser.v2i1.229.