Jurnal Sistem Informasi Bisnis 02(2025 )
Copyright © , JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587
On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71815

# Resolving Data Imbalance using SMOTE for the Analysis and Prediction of Hate Speech Sentences

Sutikman[a,*], Heri Sutanto[b], Aris Puji Widodo[c]

[a]Doctor of Information System, School of Post Graduate Studies, Diponegoro University, Jl. Imam Bardjo S.H., No. 5, Pleburan, Semarang, Indonesia 50241
[b]Physics Department, Faculty of Science and Mathematics, Diponegoro University, Jl. Prof. Soedarto, S.H., Tembalang, Semarang, Indonesia 50275
[c]Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Jl. Prof. Soedarto, S.H., Tembalang, Semarang, Indonesia 50275

*Submitted: February 25th, 2025; Revised: March 27th, 2025; Accepted: April 9th, 2025; Available Online: May 31st, 2025*
*DOI: 10.14710/vol15iss2pp198-203*

**Abstract**

Hate speech is characterized as a form of communication that expresses hostility or discontent towards particular individuals, groups, or ethnicities, with the intent to belittle one party. This research aims to examine hate speech expressions on Twitter, assessing their categorization as hate speech through the application of machine learning methodologies. The study incorporates feature engineering techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and the Synthetic Minority Over-sampling Technique (SMOTE), to mitigate challenges related to data imbalance. The machine learning models utilized include Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), and Random Forest (RF). Among these models, Logistic Regression (LR) demonstrated the highest efficacy, achieving an accuracy of 91.43%, precision of 88.83%, recall of 93.99%, and an F1 score of 97.10%.97.10%.

***Keywords:*** Hate speech; Machine learning; Text classification; TF-IDF; Smote; Text Mining.

## 1. Introduction

Social media is an essential communication tool in Indonesia, facilitating the exchange of ideas and discussions on important topic. Twitter, in particular, is widely used for public expression, allowing users to share opinions freely through tweets. These tweets often become trending topics, especially when addressing controversial issues, leading to diverse public reactions. Responses can be positive, negative, or neutral. Positive responses include praise and support, while negative responses feature sarcasm, insults, and hate. Neutral responses provide general statements that confirm or refute events without clear endorsement or criticism of the original message

The messages conveyed are often associated with criminal acts, including cases of hate speech. Hate speech refers to expressions that communicate disappointment or animosity towards individuals, groups, or races, with the intent to demean the targeted party. It is imperative to manage hate speech on social media promptly through appropriate strategies to prevent the spread of misinformation, division, defamation, and other related issues. The involvement of governmental bodies, such as the Ministry of Communication and Information, along with relevant authorities, is essential in addressing the challenges posed by hate speech (Papel et al. 2024) (Ro 1999).

*) Corresponding author: sutikman@students.undip.ac.id

Previous studies, such as the work by Sindhu Abro et al. titled "Automatic Hate Speech Detection using Machine Learning: A Comparative Study," have employed feature engineering models like TF-IDF, Word2vec, and Doc2vec with algorithms including KNN, Logistic Regression, Decision Tree, Multinomial Naïve Bayes, and SVM. Using Twitter data labeled via CrowdFlower, the study categorized tweets into hate speech, not offensive, and offensive but not hate speech, achieving 79% accuracy with SVM (Abro et al. 2020).

The study by Aditya Perwira Joan Dwitama, Dhomas Hatta Fudholi, and Syarif Hidayat, titled "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)," compared LSTM, Bi-LSTM, and CNN algorithms. The results showed that Bi-LSTM with a single layer achieved the highest accuracy of 97.66%, using the IndoBERT model for feature extraction(Patihullah and Winarko 2019).

The objective of this research is to perform an analysis focused on predicting hate speech sentences through the implementation of machine learning models. During the feature engineering process, Term Frequency — Inverse Document Frequency (TF-IDF) (Patihullah and Winarko 2019) is used to assign weights to each term. To counteract the imbalance in data distribution, the Synthetic Minority Over-sampling Technique (SMOTE) is applied (Ahammed

et al. 2020). To ensure a balanced data distribution, traditional machine learning algorithms, such as Logistic Regression (LR)(Perwira Joan Dwitama, Hatta Fudholi, and Hidayat 2023) and Decision Tree (DT)(Patihullah and Winarko 2019), were utilized. Additionally, ensemble methods, specifically Gradient Boosting (GB)(Candanedo, Feldheim, and Deramaix 2017) and Random Forest (RF)(Ro 1999), were formulated to further enhance model performance.

## 2. Research Method

This research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of six key stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These stages are essential for developing a data mining process to assess hate speech statements. The overall methodology is illustrated in Figure 1 and will be detailed in the following sections.

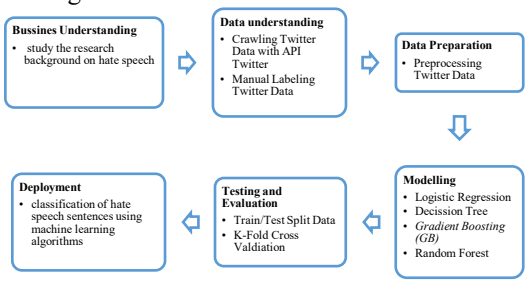


Figure 1. Research Methodology.

### *2.1. Business Understanding*

Hate speech is characterized by messages that express disappointment or hatred towards specific individuals, groups, or races, intended to belittle the targeted entity (Schröer, Kruse, and Gómez 2021). This study aims to analyze and predict hate speech on Twitter, categorizing sentences as hate speech or non-hate speech. Data will be collected via API, followed by data cleansing, preparation, and vectorization through feature engineering. The models will be evaluated using traditional and ensemble machine learning algorithms with K-fold Cross Validation to determine which algorithm performs best in automatic predictions(Khanday et al. 2022).

### *2.2. Data Understanding*

The dataset used in this research, titled “The Dataset for Hate Speech Detection in Indonesia,” was gathered during the 2017 DKI Jakarta gubernatorial elections, marked by public dissent due to candidates' minority status and age differences. Twitter data was collected in two phases: February and April 2017. The data collection focused on hashtags such as #DebatPilkadaDKI and #SidangAhok, among others. A total of 40,000 tweets were gathered, and after a cleaning process and the removal of irrelevant tweets, 1,100 tweets were manually labeled as either hate speech or non-hate speech. The details of the dataset are summarized in Table 1(Ro 1999).

Table 1. Dataset of Hate Speech Manually Labeled

| No | *Tweet* | Label |
|---|---|---|
| 1 | Senang dong ada yg fans mati sampai lupa kalau dia bego banget https://t.co/ctJFC5a8Zz,"Glad, someone is dead until he forgot that he was so stupid https://t.co/ctjfc5a8zz" | *hate speech* |
| 2 | Kami lebih resah anda jd presiden!!!! hidup anda penuh drama dan pencitraan, main FTV aja joko judulnya https://t.co/T2BUgLTnUO","We are more uneasy to be the president !!!! Your life is full of drama and imaging, just play ftv joko the title https://t.co/t2bugltnuo" | *hate speech* |
| 3 | Boleh tinggal di pinggir kali meski bukan tanahnya boleh jualan di pinggir jalan,May stay on the edge of the river even though it is not the land can sell on the side of the road | *non hate speech* |
| 4 | #MataNajwaDebatJakarta Paslon 3 serang terus dengan membawa2 hal2 yg biasa menyerang, tp Paslon 2 tenang menjawabnya","#MataJWADEBATJAKARTA PASLON 3 SERANG continues | *non hate speech* |

### *2.3. Data Preparation*

Manual labeling was conducted on this Twitter dataset with the aim of maximizing the labeling outcomes, as it can incorporate the values and meanings embedded within the tweets. During the labeling process, several individuals with diverse backgrounds were involved to ensure that the resulting labels would be more objective. The labels assigned will indicate whether a tweet is classified as hate speech or not; if it is, it will be labeled as "HS," and if it is not, it will be labeled as "NonHS." The outcome of the manual labeling process resulted in 713 labeled tweets from an initial 1,100, with 260 tweets categorized as "HS" and 453 tweets categorized as "NonHS." The results indicate an imbalanced data distribution. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) may be employed (Febiana Anistya and Erwin Budi Setiawan 2021) (Alfina et al. 2017).

The subsequent process is Data Pre-Processing, which includes the following stages:

#### *2.3.1. Remove Punctuations*

The hate speech dataset will be subjected to a data purification process, which entails the exclusion of sentences that feature Uniform Resource Locators (URLs), hashtags, emoticons, mentions, and superfluous characters such as (,.;: @#, etc.) (Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa’ah 2021)

#### *2.3.2. Lowercase*

The dataset, having been purged of irrelevant characters, will be converted to lowercase through the use of the str.lower() function from Python's libraries. This adjustment is intended to equalize the weight

assigned to each word (Zaidi, Tariq, and Belhaouari 2021).

*2.3.2. Tokenizing*

This procedure involves the division of characters into multiple sentences or segments (words/phrases), which is known as tokenization.

*2.3.4. Filtering*

Sentences that lack meaning within the dataset will be removed to enhance the classification process. This research employs Indonesian stopwords obtained from the NLTK library to filter the DataFrame and incorporates kamusalay.csv, which contains colloquial language used in Indonesia (Perwira Joan Dwitama, Hatta Fudholi, and Hidayat 2023).

*2.3.5. Stemming*

The process of stemming involves the elimination of prefixes and suffixes from words to obtain their fundamental or root form. For this purpose, the Sastrawi library is employed to perform stemming in the Indonesian language (Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah 2021).

*2.4. Modeling*

In the modeling phase, predictive analysis will be conducted using traditional machine learning algorithms and ensemble methods. Models will be built and optimized to find the best configuration for the data. Evaluation of these models will be addressed in the subsequent evaluation phase using the K-Fold Cross Validation approach (Patihullah and Winarko 2019).

*2.4.1. Term Frequency – Inverse Document Frequency (TD-IDF)*

Term Frequency — Inverse Document Frequency (TF-IDF) is a method designed to calculate the weight of each word by assessing the term frequency (TF) and inverse document frequency (IDF) for every token in each document. In essence, the TF-IDF method quantifies the frequency of a word's occurrence within a document (Febiana Anistya and Erwin Budi Setiawan 2021) (Abro et al. 2020) (Alfina et al. 2017), shown in Equation 1.

$$W_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{p} n_{j,i}} log_2 \frac{D}{d_j} \quad (1)$$

where

$W_{i,j}$ : weighting of the term $j$ in the document $i$ using the TF-IDF method
$n_{i,j}$ : frequency of the $j$-th term in the $i$-th document
$p$ : number of terms that are formed
$\sum_{j=1}^{p} n_{j,i}$ : total occurrences of all terms in document $i$
$D$ : total number of documents
$d_j$ : quantity of documents that contain the term $j$-th.

*2.4.1. Logistic Regression*

The regression algorithm employed when the data type is characterized by a binary dependent variable is known as Logistic Regression. This method does not require a linear relationship between the independent and dependent variables. Logistic Regression is categorized into two types: single logistic regression, which involves a single type of input variable, and multiple logistic regression (Alfina et al. 2017)[2(Taradhita and Putra 2021)(Cao et al. 2019), as shown in Equation 2.

$$In\left(\frac{\acute{P}}{1-\acute{P}}\right) = B_0 + B_1 X \quad (2)$$

where

$In$ : the natural logarithm
$B_0 + B_1 X$ : equation referred to as OLS
$\acute{P}$ : represents logistic probability

*2.4.3. Decision Tree (DT)*

The algorithm that employs a decision tree structure in its calculations models various possibilities, thereby identifying an alternative solution to the problem (Khanday et al. 2022)[2(Taradhita and Putra 2021)(Sheng, Chen, and Tian 2018). The formula of the decision tree algorithm is shown in Equation 3.

$$E(S) = \sum_{i=1}^{C} - P_i log_2 P_i \quad (3)$$

where $S$ represents the initial condition, $i$ denotes the set of classes within $S$, such as yes and no, and $P_i$ signifies the probability of an event occurring within $S$.

*2.4.4. Decision Tree (DT)*

The Gradient Boosting Classifier is an algorithm developed from decision trees. This algorithm employs boosting techniques for its calculations and operates sequentially by incorporating previous predictors that are less accurate into the ensemble. It does so by checking for any errors encountered in the predictions (Khanday et al. 2022) (Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah 2021), as shown in Equation 4.

$$F_0(x) = \underset{\Upsilon}{\operatorname{argmin}} \sum_{i=I}^{n} L(y_1, \Upsilon) \quad (4)$$

where

$F_0$ : This represents the prediction of the initial constant value
L : This represents a quadratic loss function
argmin : The task involves determining the predictive value or gamma that is

designed to achieve the lowest possible losses

ϒ : The value of gamma will serve as the basis for making predictions

*2.4.5. Synthetic Minority Over-sampling Technique (SMOTE)*

The Synthetic Minority Over-sampling Technique (SMOTE) is a method employed to address the issue of oversampling in datasets. SMOTE does not eliminate data when the distribution is imbalanced, taking into account the impact on decision boundaries within the feature space. This technique generates additional samples for the minority class, thereby enhancing the classification process by allowing for a more comprehensive analysis of the minority class data.

*2.4.6. Random Forest*

Random Forest is a machine learning algorithm that belongs to the same family as Decision Trees. It is an ensemble method that combines multiple trees to create a new model. This algorithm relies on vector values that maintain a uniform distribution across all trees, with each tree having a specified maximum depth (Febiana Anistya and Erwin Budi Setiawan 2021)[3(Abro et al. 2020)(Alfina et al. 2017)(Ro 1999).

*2.5. Testing and Evaluation*

The model's evaluation technique utilized is K Fold Cross Validation, which entails partitioning the dataset in an 80:20 ratio, where 80% is reserved for training purposes and 20% is assigned for testing. The parameter K is established at 10, signifying that the testing process will occur 10 times, followed by the computation of the average across all 10 iterations (Patihullah and Winarko 2019)(Candanedo, Feldheim, and Deramaix 2017).

*2.6. Deployment*

At this stage, a report will be prepared detailing the implementation process of data mining, aimed at providing an overview of the conclusions drawn from the data mining algorithm calculations.

## 3. Results and Analysis

The outcomes obtained from the data pre-processing procedure, including steps such as removing punctuations, converting text to lowercase, tokenizing, filtering, and stemming, are summarized in Table 2.

Table 2. Results of Data Pre-Processing Using Python

| *Proses* | Data |
|---|---|
| Original Data | Yg no.1 nih ngomongin hal yg di luar program mulu ya. Ngga ada bhan ya bwt serang lawan.. Cape dehh! #DebatFinalDKI #DebatFinalPilkadaJKT, |
| *Remove Punctuations* | Yg no 1 nih ngomongin hal yg di luar program mulu ya Ngga ada bhan ya bwt serang lawan Cape dehh |
| *Lowercase* | yg no 1 nih ngomongin hal yg di luar program mulu ya ngga ada bhan ya bwt serang lawan cape dehh |
| *Tokenizing* | [yg [no [1 [nih [ngomongin [hal [yg [di [luar [program [mulu [ya [ngga [ada [bhan [ya [bwt [serang [lawan [cape [dehh |
| *Filtering* | [ngomongin [hal [di [luar [program [ya [ada [ya [serang [lawan |
| *Stemming* | [ngomong [hal [di [luar [program [ya [ada [ya [serang [lawan |

In the pre-processing steps mentioned earlier, Python was utilized for removing stopwords with the Indonesian language library NLTK, and stemming was carried out using the Sastrawi library. The dataset exhibited an imbalanced label distribution. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to create synthetic samples for the minority class using data from the majority class. Initially, the class labels were imbalanced at 203:367. After applying SMOTE, the ratio was balanced to 367:367, creating an equal class distribution. The bar is plotted in Figure 2. It shows the data distribution before and after SMOTE, illustrating the balanced dataset(Papel et al. 2024)(Ahammed et al. 2020).

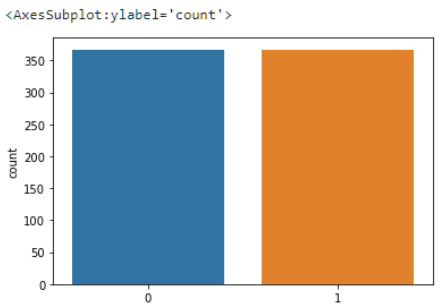


Fig 2. Distribution of data following the application of SMOTE.

The parameters in traditional machine learning algorithms and ensemble machine learning are set to null or empty values, with the aim of observing the final outcomes using the default parameters. The parameters used in this research are shown in Table 3.

Table 3. Model Algoritma Machine Learning Tradisional dan Ensemble

| | |
|---|---|
| LGR | LogisticRegression() |
| RF | RandomForestClassifier() |
| DT | DecisionTreeClassifier() |
| GB | GradientBoostingClassifier() |

On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71815

The following section presents a comparison of the results obtained with and without the application of the Synthetic Minority Over-sampling Technique (SMOTE), highlighting the values of Accuracy, Precision, Recall, and F1 score. These outcomes were computed using various machine learning algorithms and evaluated with a K-Fold Cross Validation model, where K was set to 10.

The comparison of accuracy values indicates that the Logistic Regression (LR) algorithm, when employing the SMOTE method, achieves an accuracy rate of 91.43%. In contrast, when the SMOTE method is not utilized, the accuracy of Logistic Regression (LR) decreases to 79.30%. The comparison of accuracy values across the algorithms is illustrated in Figure 3.

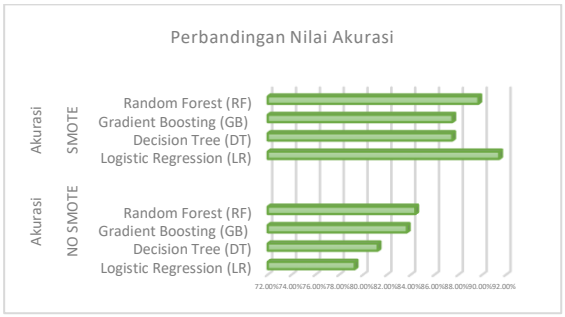


Fig 3. Comparison of Accuracy Values with and without the application of the SMOTE method.

The comparison of accuracy values indicates that the Logistic Regression (LR) algorithm, when employing the SMOTE method, achieves a Precision score of 88.83%. In contrast, when the SMOTE method is not utilized, the accuracy of Logistic Regression (LR) declines to 45.33%. The comparison of accuracy values across the algorithms is illustrated in Figure 4.

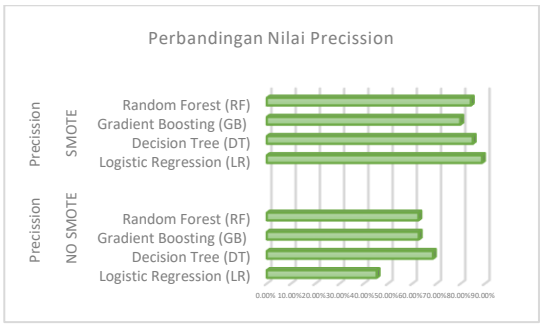


Fig 4. Comparison of Precision Values With and Without the Use of the SMOTE Method.

The comparison of accuracy values indicates that the Random Forest (RF) algorithm, when employing the SMOTE method, achieves a Recall value of 96.17%. In contrast, when the SMOTE method is not utilized, the accuracy of the Random Forest (RF) algorithm decreases to 91.75%. The comparison of accuracy values across the algorithms is illustrated in Figure 5.

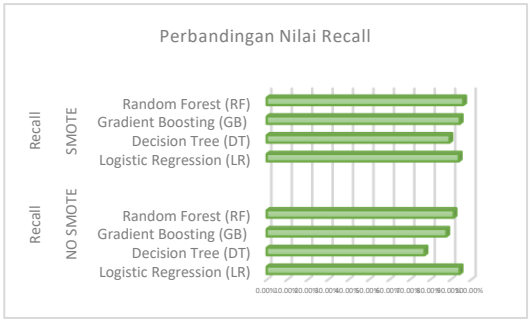


Fig 5. Comparison of Recall Values with and without the application of the SMOTE method.

Finally, the comparison of accuracy values indicates that the Logistic Regression (LR) algorithm, when employing the SMOTE method, achieves an F1 score of 97.10%. In contrast, when the SMOTE method is not utilized, the accuracy of Logistic Regression (LR) decreases to 92.70%. The comparison of accuracy values across the algorithms is illustrated in Figure 6.

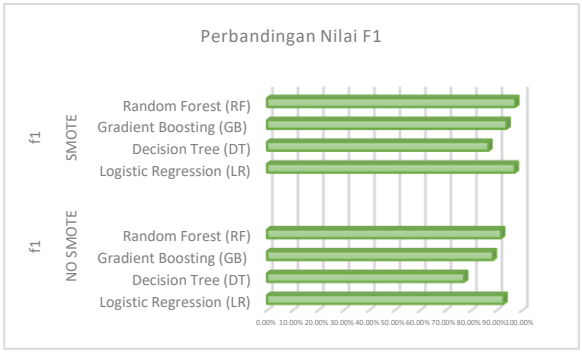


Fig 6. Comparison of F1 Scores with and without the application of the SMOTE method.

## 4. Conclusion

The study concludes that the prediction of hate speech sentences can be effectively achieved by categorizing them into two labels: hate speech and not hate speech, utilizing four machine learning algorithms Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), and Random Forest (RF). The implementation of the Synthetic Minority Over-sampling Technique (SMOTE) plays a crucial role in enhancing the performance metrics, such as accuracy, precision, recall, and F1 score, thereby effectively addressing data imbalance issues. A comparative assessment of the algorithms indicates that SMOTE significantly improves the performance of all tested models, confirming its value in predictive modeling of hate speech. Among the algorithms evaluated, Logistic Regression (LR) is identified as the most effective, achieving the highest accuracy of 91.43%, precision of 88.83%, recall of 93.99%, and an

F1 score of 97.10%, making it the best performer in this study.

**References**

Abro, Sindhu, Sarang Shaikh, Zafar Ali, Sajid Khan, Ghulam Mujtaba, and Zahid Hussain Khand. 2020. "Automatic Hate Speech Detection Using Machine Learning: A Comparative Study." *International Journal of Advanced Computer Science and Applications* 11 (8): 484–91. https://doi.org/10.14569/IJACSA.2020.0110861.

Ahammed, Khair, Md Shahriare Satu, Md Imran Khan, and Md Whaiduzzaman. 2020. "Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods." *2020 IEEE Region 10 Symposium, TENSYMP 2020*, no. June: 1371–74. https://doi.org/10.1109/TENSYMP50017.2020.9230464.

Alfina, Ika, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study." *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017* 2018-January (October): 233–37. https://doi.org/10.1109/ICACSIS.2017.8355039.

Candanedo, Luis M., Véronique Feldheim, and Dominique Deramaix. 2017. "Data Driven Prediction Models of Energy Use of Appliances in a Low-Energy House." *Energy and Buildings* 140: 81–97. https://doi.org/10.1016/j.enbuild.2017.01.083.

Cao, Guogang, Mengxue Li, Cong Cao, Ziyi Wang, Meng Fang, and Chunfang Gao. 2019. "Primary Liver Cancer Early Screening Based on Gradient Boosting Decision Tree and Support Vector Machine." *ICIIBMS 2019 - 4th International Conference on Intelligent Informatics and Biomedical Sciences*, 287–90. https://doi.org/10.1109/ICIIBMS46890.2019.8991441.

Febiana Anistya, and Erwin Budi Setiawan. 2021. "Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5 (6): 1044–51. https://doi.org/10.29207/resti.v5i6.3521.

Khanday, Akib Mohi Ud Din, Syed Tanzeel Rabani, Qamar Rayees Khan, and Showkat Hassan Malik. 2022. "Detecting Twitter Hate Speech in COVID-19 Era Using Machine Learning and Ensemble Learning Techniques." *International Journal of Information Management Data Insights* 2 (2): 100120. https://doi.org/10.1016/j.jjimei.2022.100120.

Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah. 2021. "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network Dan Naïve Bayes." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5 (4): 802–8. https://doi.org/10.29207/resti.v5i4.3308.

Papel, Habibur Rahman, Udoy Chandra Dey, Toufiq Hasan Turza, Avijit Datta, and Tanu Sarkar. 2024. "Bangla Hate Speech Detection by Embedding and Hybrid Machine Learning Algorithms."

Patihullah, Junanda, and Edi Winarko. 2019. "Hate Speech Detection for Indonesia Tweets Using Word Embedding And Gated Recurrent Unit." *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 13 (1): 43. https://doi.org/10.22146/ijccs.40125.

Perwira Joan Dwitama, Aditya, Dhomas Hatta Fudholi, and Syarif Hidayat. 2023. "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 7 (2): 302–9. https://doi.org/10.29207/resti.v7i2.4642.

Ro, Sex. 1999. "Greek a d Roman MytholoGy," 473–81. https://support.twitter.eom/articles/l.

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez. 2021. "A Systematic Literature Review on Applying CRISP-DM Process Model." *Procedia Computer Science* 181 (2019): 526–34. https://doi.org/10.1016/j.procs.2021.01.199.

Sheng, Peng, Li Chen, and Jing Tian. 2018. "Learning-Based Road Crack Detection Using Gradient Boost Decision Tree." *Proceedings of the 13th IEEE Conference on Industrial Electronics and Applications, ICIEA 2018*, 1228–32. https://doi.org/10.1109/ICIEA.2018.8397897.

Taradhita, Dewa Ayu Nadia, and I. Ketut Gede Darma Putra. 2021. "Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network." *Journal of ICT Research and Applications* 14 (3): 225–39. https://doi.org/10.5614/itbj.ict.res.appl.2021.14.3.2.

Zaidi, Syed Ali Jafar, Saad Tariq, and Samir Brahim Belhaouari. 2021. "Future Prediction of Covid-19 Vaccine Trends Using a Voting Classifier." *Data* 6 (11). https://doi.org/10.3390/data6110112.