

# Modeling sentiment analysis of Indonesian biodiversity policy Tweets using IndoBERTweet

Mohammad Teduh Uliniansyah, Asril Jarin, Agung Santosa, Gunarso

Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics,  
National Research and Innovation Agency (BRIN), KST Samaun Samadikun Bandung, Bandung, Indonesia

## Article Info

### Article history:

Received Aug 28, 2024

Revised Feb 12, 2025

Accepted Mar 15, 2025

### Keywords:

BERT embeddings

Biodiversity policy

IndoBERTweet

Sentiment analysis

Twitter data

## ABSTRACT

This study develops and evaluates a sentiment analysis model using IndoBERTweet to analyze Twitter data on Indonesia's biodiversity policy. Twitter data focusing on topics such as food security, health, and environmental management were collected, with a representative subset of 13,435 tweets annotated from a larger dataset of 500,000 to ensure reliable sentiment labels through majority voting. IndoBERTweet was compared to seven traditional machine-learning classifiers using TF-IDF and BERT embeddings for feature extraction. Model performance was assessed using mean accuracy, mean F1 score, and statistical significance (p-values). Additionally, sentiment analysis included word attribution techniques with BERT embeddings, enhancing relevance, interpretability, and consistent attribution to deliver accurate insights. IndoBERTweet models consistently outperformed traditional methods in both accuracy and F1 score. While BERT embeddings boosted performance for conventional models, IndoBERTweet delivered superior results, with p-values below 0.05 confirming statistical significance. This approach demonstrates that the model's outputs are explainable and align with human understanding. Findings underscore IndoBERTweet's substantial impact on advancing sentiment analysis technology, showcasing its potential to drive innovation and elevate practices in the field.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Asril Jarin

Research Center for Data and Information Science, National Research and Innovation Agency

Bandung, Indonesia

Email: asri003@brin.go.id

## 1. INTRODUCTION

Indonesia, recognized as one of the world's most significant biodiversity hotspots with a high concentration of endemic species [1], is facing increasing threats from habitat destruction, climate change, and environmental degradation. These challenges underscore the urgent need for effective conservation policies, which must be informed by public sentiment to ensure their responsiveness and effectiveness. However, analyzing public sentiment from large-scale social media platforms like Twitter presents significant challenges due to the informal, diverse nature of the language used. To address these challenges, this study utilizes IndoBERTweet [2], a pre-trained language model optimized for Indonesian Twitter data, to perform sentiment analysis on the public discourse surrounding Indonesia's biodiversity policies. By leveraging context-rich embeddings, IndoBERTweet enhances both the accuracy and interpretability of sentiment analysis [3], offering a more effective tool than traditional machine learning approaches. This study aims to provide actionable insights to policymakers by applying these advanced methods to refine biodiversity conservation strategies and improve public engagement [4].

Building on the demonstrated strengths of IndoBERTweet, this study further evaluates the performance of its different variants (GELU, Tanh, and None) as described by Santosa *et al.* [5], comparing them with traditional machine learning classifiers such as logistic regression, support vector machine, and random forest. Several studies have highlighted the effectiveness of BERT and its variants in tweet sentiment analysis, demonstrating significant gains in accuracy and contextual understanding when combined with neural network architectures. For instance, Bello *et al.* [6] proposed a BERT framework for sentiment analysis that integrates BERT with convolutional neural network (CNN), recurrent neural network (RNN), and bidirectional long short-term memory (BiLSTM), achieving improvements in accuracy, precision, and recall compared to conventional methods. These traditional models, which typically rely on feature extraction methods like TF-IDF [7], often fall short in handling the complexity and informal language typical of social media. In contrast, IndoBERTweet, leveraging BERT embeddings [8], provides a more nuanced analysis. This study enhances prior evaluations by integrating advanced techniques such as word attribution and ten-fold cross-validation, using metrics like mean accuracy and mean F1-score, with statistical significance confirmed through p-values. These comprehensive evaluations demonstrate the superior performance of IndoBERTweet and provide more reliable tools for policy decision-making.

The architecture of the IndoBERTweet variants, as shown in Figure 1, illustrates a deep neural network (DNN) structure beginning with a pre-trained IndoBERTweet model designed to capture contextual embeddings and semantic representations unique to Bahasa Indonesia tweets. This model is then followed by two fully connected layers: a 256-unit layer that matches the IndoBERTweet pooled output vector and a three-unit output layer representing sentiment classification outcomes. Three different activation functions—None, GELU, and Tanh—are tested in the first fully connected layer to assess their impact on sentiment analysis performance. This study further demonstrates that IndoBERTweet surpasses seven traditional classifiers, which rely on methods like TF-IDF and BERT embeddings, in analyzing tweets about Indonesia's biodiversity policies. With a focused dataset of 13,435 tweets, IndoBERTweet achieves superior mean accuracy and F1 scores, with statistically significant p-values (below 0.05) compared to conventional models. Additionally, word attribution techniques enhance interpretability, supporting the model's application in high-precision policy-making for biodiversity and environmental management.

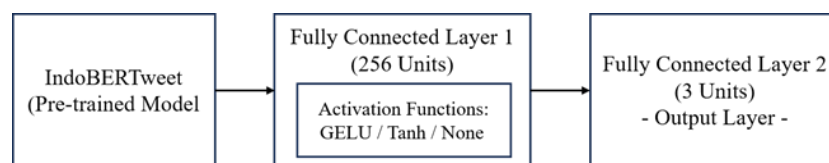


Figure 1. DNN architecture of the IndoBERTweet model with two fully connected layers [5]

The paper is organized as follows: in the method section, we describe the process of collecting and preparing Twitter data on biodiversity policies in Indonesia, followed by the development and evaluation of the sentiment analysis models using IndoBERTweet and traditional classifiers. The results and discussion section presents a detailed comparison of the performance of IndoBERTweet against traditional models, supported by word-attribution techniques and statistical analysis. Finally, the conclusion summarizes the key findings, highlighting the practical implications for biodiversity policy-making and offering directions for future research to enhance sentiment analysis methodologies in similar contexts.

## 2. METHOD

As illustrated in Figure 2, the diagram presents the steps for developing and evaluating the sentiment analysis model using IndoBERTweet on Indonesian Twitter data about biodiversity policy. The process begins with data preparation, which includes collecting, cleaning, selecting, and labeling the data from Twitter. Next, data preprocessing involves normalizing the data and ensuring consistent data formats. The modeling phase involves both deep learning and non-deep learning techniques. For deep learning, the IndoBERTweet model is utilized with various activation functions. Feature vectorization is performed using TF-IDF and BERT embeddings for non-deep learning techniques. Finally, model performance evaluation uses 10-fold cross-validation, measuring accuracy and F1-score, assessing statistical significance with p-values, and evaluating word attribution. This comprehensive approach ensures a robust assessment of the model's ability to analyze sentiment in the context of Indonesian biodiversity policy.

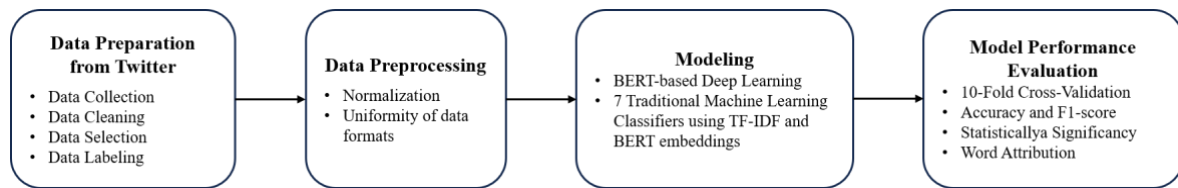


Figure 2. Workflow of sentiment analysis model development and evaluation using IndoBERTweet on Indonesian Twitter data

## 2.1. Dataset preparation from Twitter

The dataset preparation phase involved creating a labeled dataset of tweets from Twitter, categorizing them as positive, negative, or neutral for sentiment analysis. Tweets were collected using the Twitter API [9] from January 2020 to March 2023, focusing on topics related to food security, healthcare, and environmental management in Indonesia. This process initially gathered 500,000 tweets. After cleaning to remove duplicates, non-Indonesian content, and irrelevant information such as advertisements and novel quotes, the dataset was reduced to 200,015 tweets. Due to limitations in annotator availability, 15,323 tweets were selected for annotation. To ensure consistency and reliability in the annotation process, 18 trained annotators were provided with detailed guidelines for sentiment classification based on established standards [4] from prior studies. Each tweet was reviewed independently by three annotators, and inter-annotator agreement was assessed using Fleiss' Kappa with a value of 0.62187, which showed a substantial agreement. A majority voting approach was employed to finalize the sentiment label for each tweet, resolving any disagreements. This approach ensured that the annotations were consistent and reflected a reliable consensus among the annotators. After excluding tweets labeled as "none," the final dataset used in the research comprised 13,435 tweets. Detailed information on the labeling process and the final dataset is provided in the related paper [10], and the dataset is available at the Mendeley Data repository [11].

## 2.2. Data preprocessing

Following the initial dataset preparation, several preprocessing steps were performed to ready the data for modeling. In line with the approach by Pebiana *et al.* [12], the text was transformed to lowercase to ensure uniformity. URLs were removed, punctuation (except apostrophes) was replaced with spaces, and non-ASCII characters were substituted with their nearest ASCII equivalents. Informal language and typographical errors were normalized, word variations were standardized, and numeric data and non-ASCII characters were eliminated. Consecutive spaces were consolidated, and stopwords were removed using the Sastrawi Python library [13], though adverbs were retained due to their significant role in sentiment analysis. These preprocessing steps prepared the dataset for modeling by ensuring cleaner and more uniform text data.

One of the major challenges encountered during the study was the informal and diverse nature of the language used in Indonesian Twitter, which includes slang, abbreviations, and mixed languages. This made it difficult to clean and preprocess the data effectively. Additionally, obtaining reliable sentiment labels required careful curation and annotation of tweets, as well as resolving disagreements between annotators, which added complexity to the dataset preparation process.

## 2.3. Modeling

This study compares sentiment analysis models by employing both deep learning and traditional machine learning techniques, each offering distinct advantages for text classification. The deep learning approach uses a BERT architecture, specifically *IndoBERT*, a model pre-trained on Indonesian Twitter data and thus highly effective at handling informal language commonly found on social media. *IndoBERT*'s contextual embeddings provide nuanced semantic representations that make it particularly suited to sentiment analysis in Indonesian Twitter data, where language patterns can be complex and context-sensitive.

The traditional machine learning methods selected include logistic regression, support vector classifier, random forest, LGBMClassifier, extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), and decision tree, chosen for their proven effectiveness in text classification and sentiment analysis tasks. These models offer a balance of interpretability, robustness, and performance: logistic regression and support vector classifiers are straightforward and efficient, random forest and decision tree add ensemble stability to reduce overfitting, and gradient boosting methods like LGBMClassifier and XGBoost provide high accuracy for complex datasets. While simpler models such as naive Bayes were considered, they were excluded after initial tests indicated limited performance on this context-heavy Twitter dataset. To ensure a thorough comparison, we employed both TF-IDF and BERT embeddings for feature representation in all traditional models, allowing a direct evaluation of traditional vectorization versus BERT-based contextual

embeddings. Results are summarized in tables in the results and discussion section, with key performance metrics, including mean accuracy and F1-score, presented concisely for clarity.

### 2.3.1. Modeling sentiment analysis with BERT-based deep learning techniques

The pre-trained IndoBERTweet model [2] has proven highly effective for sentiment analysis on Indonesian Twitter data due to its specific training in the informal language typical of social media. Unlike IndoBERT [3], which was trained on over 220 million words from various formal Indonesian sources like Wikipedia and news articles, IndoBERTweet's training corpus consists of approximately 409 million word tokens solely from Indonesian Twitter data. This focus on informal language enables IndoBERTweet to better capture the nuances of Indonesian social media discourse, making it highly suitable for analyzing sentiments on Twitter. The model's training employs unsupervised learning techniques, predicting masked tokens in input text, consistent with BERT's architecture, which enhances its ability to understand contextual relationships within social media text.

Building upon IndoBERTweet's strengths, our sentiment analysis system incorporates two fully connected layers that utilize sentence embeddings generated by IndoBERTweet as input features. This architecture, illustrated in Figure 1, is structured as a DNN to capture complex contextual relationships inherent in Indonesian language text, essential for accurate sentiment analysis. Additionally, we evaluated the performance impact of different activation functions—specifically, hyperbolic tangent (tanh), Gaussian error linear unit (GELU), and no activation—in the first fully connected layer to optimize the model's effectiveness. These comparisons provide insights into how IndoBERTweet embeddings interact within a deep learning framework, contributing to more informed choices in model configurations for improved sentiment analysis outcomes.

### 2.3.2. Modeling sentiment analysis with traditional machine learning methods

In this section, we investigate the effectiveness of seven traditional machine learning methods in sentiment analysis by utilizing two distinct vectorization techniques: TF-IDF and BERT embeddings. While TF-IDF provides a statistical measure of word importance within the corpus, BERT embeddings offer rich, contextualized word representations that capture semantic nuances. The traditional machine learning methods we employ are logistic regression, support vector machine, random forest, LGBMClassifier, XGBoost, AdaBoost, and decision tree. To thoroughly assess the performance and precision of these methods in sentiment analysis, we divide our investigation into two approaches: TF-IDF-based traditional models and BERT embedding-based traditional models. This analysis provides insights into the strengths and limitations of traditional techniques in capturing sentiment from textual data, compared with BERT-based deep learning approaches.

- TF-IDF-based traditional models: the TF-IDF technique [7] is an essential tool for feature extraction from the labeled dataset, enabling the application of various traditional machine-learning algorithms. This approach begins with logistic regression [14], a method that models the probabilities of binary outcomes. Next, we utilize the support vector classifier [15], which effectively employs hyperplanes to separate classes in high-dimensional space. Light gradient boosting machine [16] is also applied, and it is known for its high efficiency when handling large-scale datasets. Additionally, random forest [17] is employed as an ensemble method that constructs multiple decision trees to improve model accuracy. XGBoost [18], optimized for both performance and efficiency, is included in our analysis. We further incorporate AdaBoost [19], which iteratively combines weak classifiers to form a robust classifier. Finally, a decision tree [20], with its tree-like structure, is used for both classification and regression tasks. This comprehensive evaluation allows for a detailed comparison of how these traditional machine learning models perform in the context of sentiment analysis when using TF-IDF as the feature representation technique.
- BERT embedding-based traditional models: BERT embeddings provide a significant advantage over TF-IDF vectors by capturing the context in which words appear, creating a more holistic representation of text. TF-IDF, a statistical metric for evaluating a word's importance within a corpus, effectively identifies keyword relevance but treats each term independently, lacking contextual insight. In contrast, BERT, or bidirectional encoder representations from transformers, generates contextualized word embeddings that capture complex semantic and syntactic nuances. This capability is particularly valuable in sentiment analysis, where a word's sentiment can vary substantially based on nearby terms. Research has shown that using BERT embeddings as input features for traditional, non-deep-learning classifiers—such as support vector machines, random forests, or boosting algorithms—enables these models to harness the enriched, contextual information embedded in the text, leading to improved accuracy and robustness in sentiment classification tasks compared to models relying solely on TF-IDF vectors. Employing BERT embeddings in this way offers a promising approach to enhancing sentiment analysis models [21].

## 2.4. Model performance evaluation

This study addresses the challenge of analyzing public sentiment from the informal and diverse language used on Indonesian Twitter, particularly in the context of biodiversity policies, by evaluating the effectiveness of IndoBERTweet compared to seven traditional classifiers. Using a curated dataset of 13,435 tweets, we employ both TF-IDF and BERT embeddings for feature extraction in the traditional classifiers, while IndoBERTweet leverages its pre-trained embeddings. Key metrics such as mean accuracy, mean F1-score, and statistical significance (p-values) are used to compare the models to ensure a robust performance assessment. Further evaluation methods include 10-fold cross-validation to ensure robustness, accuracy, and F1-score to measure predictive correctness, Statistical Significance tests to validate performance differences, and Word Attribution to enhance interpretability by identifying key terms influencing sentiment predictions. These evaluations provide critical insights into the strengths and limitations of each approach for analyzing sentiment from Indonesian Twitter data.

### 2.4.1. 10-fold cross-validation

10-fold cross-validation is a widely recognized method for evaluating machine learning models [22]. This technique divides the dataset  $D$  into ten equal subsets, known as “folds.” In each iteration, one-fold  $F_k$  is set aside as the testing set, while the remaining nine folds (90% of the data) are used for training the model (10% for testing). This process is repeated ten times, with each subset serving as the test set exactly once. The performance metric  $M_k$  is calculated for each fold, and the overall performance is determined by averaging these metrics across all ten folds (1).

$$\text{Mean Performance} = \frac{1}{10} \sum_{k=1}^{10} M_k \quad (1)$$

This average offers a comprehensive assessment of the model’s capability to generalize to unseen data.

10-fold cross-validation enhances evaluation reliability by reducing variance from random data splits and ensuring each instance has equal chances in training and testing sets, reducing overfitting risks. Scikit-learn’s `cross_val_score` function automates data division, model training, and performance evaluation. Performance metrics are averaged across all folds for a robust estimate, making this method especially effective when data is limited and a single train-test split may not accurately reflect model generalization.

### 2.4.2. Accuracy and F1-score

Building on the 10-fold cross-validation process, we now assess model performance using accuracy and F1-score, with results averaged across all ten folds to ensure a comprehensive evaluation. Accuracy, calculated as (2), provides an overall view of model performance, where true positives (TP) and true negatives (TN) represent correctly predicted positive and negative instances, and false positives (FP) and false negatives (FN) account for incorrect predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

While accuracy gives a broad overview, the F1-score offers a more nuanced evaluation, particularly for imbalanced datasets. Precision, the proportion of TP out of all positive predictions (TP + FP), reflects the model’s ability to avoid false positives, while recall, the proportion of TP out of all actual positives (TP + FN), indicates the model’s effectiveness in identifying true positives. The F1-score, defined as the harmonic mean of precision and recall, is given by (3).

$$F1\_score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

This metric balances the trade-off between precision and recall, making it especially useful when positive and negative classes are not equally represented. Calculating both accuracy and F1-score for each fold during cross-validation ensures that our evaluation reflects the model’s performance across different data subsets, enhancing the robustness of our results.

### 2.4.3. Statistical significance

To assess the reliability of performance differences among the sentiment analysis models, we use the paired t-test [23]. This statistical test compares the mean accuracy and F1-score of the models, which include BERT-based deep learning and traditional machine learning models employing either TF-IDF or BERT embeddings. The paired t-test evaluates whether the observed differences in performance metrics across the 10-fold cross-validation are statistically significant or if they could be attributed to random variation [24].

The t-value for the paired t-test is calculated as (4), where  $\bar{d}$  is the mean of the differences between paired observations,  $s_d$  is the standard deviation of these differences, and  $n$  is the number of folds. This t-value measures how significantly the mean difference deviates from zero, with the corresponding p-value derived from the t-distribution with  $n-1$  degrees of freedom.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (4)$$

A p-value less than 0.05 indicates that the performance differences are statistically significant [25], [26], suggesting that one model performs better than another meaningfully. This analysis ensures that our conclusions about model effectiveness are supported by statistical evidence, providing a reliable basis for evaluating the advantages of different sentiment analysis techniques [27].

#### 2.4.4. Word attribution

In addition to assessing model performance with accuracy and F1-score, this study employs word attribution techniques to deepen insights into sentiment analysis models, particularly those using BERT embeddings [28], [29]. Word attribution is crucial for understanding the influence of individual words or tokens on model predictions, ensuring that the derived insights are accurate and interpretable. To quantify each word's contribution to sentiment prediction, we use the integrated gradients method, which computes the average gradient of the model's output with respect to word embeddings integrated from a baseline (e.g., a zero vector) to the actual input. The formula for Integrated Gradients is defined in (5).

$$IntegratedGrad_i(x) = (x_i - x'_i) \times \int_0^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (5)$$

where  $x$  is the input feature,  $x'$  is the baseline input,  $F$  is the model function, and  $\alpha$  is a scaling factor [29].

To validate and visualize results, we use word clouds to display word prominence based on attribution scores [30], [31]. Highlighting words that significantly impact sentiment predictions. This combination of quantitative attribution and qualitative visualization enhances the interpretability and reliability of sentiment analysis models.

### 3. RESULTS AND DISCUSSION

This section presents the results of comparing IndoBERTweet with traditional machine learning models to the curated dataset, with performance measured by accuracy and F1-score. We also analyze sentiment distribution across biodiversity policy subdomains and evaluate model performance through statistical significance and word-attribution analysis. These findings provide insights into the effectiveness of the models and offer recommendations for future improvements.

#### 3.1. Labeling results

The analysis of sentiment distribution within our labeled dataset of 13,435 tweets reveals distinct patterns across various subdomains of biodiversity policies in Indonesia. The health subdomain exhibits the highest proportion of positive tweets, while food security is predominantly associated with negative sentiment. These findings underscore the heterogeneous nature of public sentiment, with Figure 3 providing a comprehensive overview of positive, negative, and neutral sentiments across the subdomains.

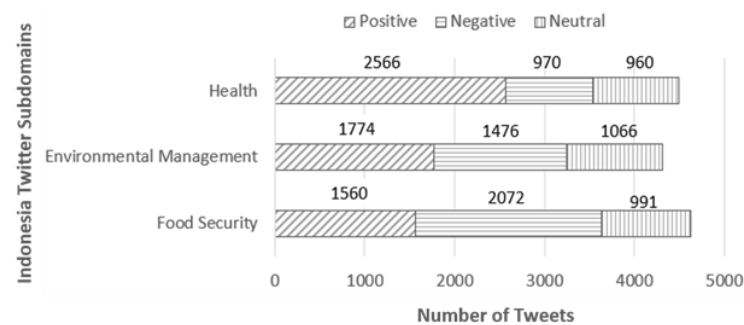


Figure 3. Distribution of sentiment labels across Indonesian Twitter subdomains

### 3.2. Evaluation results

This experiment evaluated seven non-deep-learning methods-logistic regression, support vector classifier, random forest, LGBMClassifier, XGBoost, AdaBoost, and decision tree-using both TF-IDF and BERT embeddings as input features. Additionally, the IndoBERTweet pre-trained model was tested with three variations of a fully connected layer: no activation function, GELU, and Tanh. In total, 17 sentiment analysis models were evaluated using 10-fold cross-validation, with accuracy and F1-score metrics compared across models. Statistical significance tests were conducted to compare model pairs, and the best-performing BERT-based model underwent word-attribution analysis to understand the impact of specific words on sentiment prediction.

#### 3.2.1. Performance metric

Table 1 presents the evaluation metrics for each model, including mean accuracy and mean F1 score, averaged over ten folds and highlighting the best performance from the folds. The BERT-based models demonstrated superior performance to traditional machine learning models, as illustrated in Figure 4, which shows each classifier's mean accuracy and mean F1 score. The highest mean F1 score of 0.7633 was achieved using the Tanh activation function, while the GELU activation function yielded the highest mean accuracy of 0.7899. These results underscore BERT's exceptional capability to capture the nuanced sentiment expressed within the dataset. Furthermore, the BERT-based models showed remarkable consistency in their performance, with slight variations in accuracy and F1 scores across different activation functions.

Table 1. The accuracy and F1-score for each model

Vectorization	Classifier	Mean accuracy	Std. Dev accuracy	Mean F1 score	Std. Dev F1 score	Best accuracy	Best F1 score
TF-IDF	Logistic regression	0.7144	0.0135	0.6688	0.0094	0.7337	0.6819
	Support vector classifier	0.7101	0.0142	0.6556	0.0097	0.7307	0.6689
	Random forest	0.6670	0.0137	0.6022	0.0114	0.6884	0.6212
	LGBMClassifier	0.6915	0.0118	0.6475	0.0104	0.7149	0.6700
	XGBoost	0.6762	0.0151	0.6281	0.0151	0.7119	0.6637
	AdaBoost	0.6153	0.0144	0.5688	0.0152	0.6399	0.5891
	Decision tree	0.5596	0.0100	0.5263	0.0124	0.5781	0.5479
BERT embedding	Logistic regression	0.7409	0.0087	0.7057	0.0058	0.7559	0.7145
	Support vector classifier	0.7616	0.0096	0.7198	0.0047	0.7792	0.7277
	Random forest	0.6828	0.0155	0.6075	0.0121	0.7041	0.6192
	LGBMClassifier	0.7362	0.0124	0.6952	0.0086	0.7511	0.7102
	XGBoost	0.7315	0.0122	0.6878	0.0118	0.7443	0.7040
	AdaBoost	0.6689	0.0125	0.6144	0.0120	0.6920	0.6287
	Decision tree	0.5396	0.0121	0.5053	0.0108	0.5615	0.5201
BERT	IndoBERTweet - GELU	0.7899	0.0123	0.7632	0.0140	0.8065	0.7810
	IndoBERTweet - TANH	0.7891	0.0129	0.7633	0.0156	0.8088	0.7907
	IndoBERTweet - NONE	0.7868	0.0103	0.7582	0.0153	0.8080	0.7917

Using BERT word embeddings as features for traditional machine learning models, other than Decision Tree, increased the performance of the resulting models. However, they were still lower than the BERT-based models. The performance decrease of the Decision Tree when using BERT embeddings can be attributed to the dense nature of BERT embeddings, which makes it harder to find optimal splits for the Decision Tree. In contrast, TF-IDF features are sparse, which is more suitable for Decision Trees as they can effectively leverage the sparsity for better splits.

Prior sentiment analysis studies in this field have used general machine learning and NLP models, but they have not fully addressed the unique linguistic and cultural characteristics of Indonesian social media discourse. IndoBERTweet effectively fills this research gap, demonstrating higher accuracy and F1 scores than traditional models like logistic regression, support vector machines, and random forest. Specifically, IndoBERTweet achieved a mean accuracy of 78.99% and an F1 score of 0.7633, surpassing the highest accuracy and F1 score of 71.44% and 0.6688, respectively, achieved by traditional models. Its pre-trained capabilities enable a more nuanced understanding of complex sentiment expressions specific to Indonesian biodiversity policy. While traditional models performed reasonably with TF-IDF vectorization, they struggled with the informal language prevalent on social media. Despite IndoBERTweet's advantages, traditional models remain valuable in resource-limited settings for their simpler implementation and lower computational demands. These findings underscore the importance of selecting models suited to the unique requirements of a given task and context, revealing IndoBERTweet's potential to advance more equitable and precise sentiment analysis in Indonesian biodiversity discourse.



However, this study is limited by its reliance on a single dataset from Twitter, which, while effective in evaluating IndoBERTweet’s strengths, may not fully capture sentiment nuances across other social media platforms. This limitation could affect the model’s applicability in broader contexts, as sentiment expressions may vary significantly across different platforms. As a result, the findings should be interpreted with caution, recognizing that additional datasets may be needed to confirm the model’s generalizability.

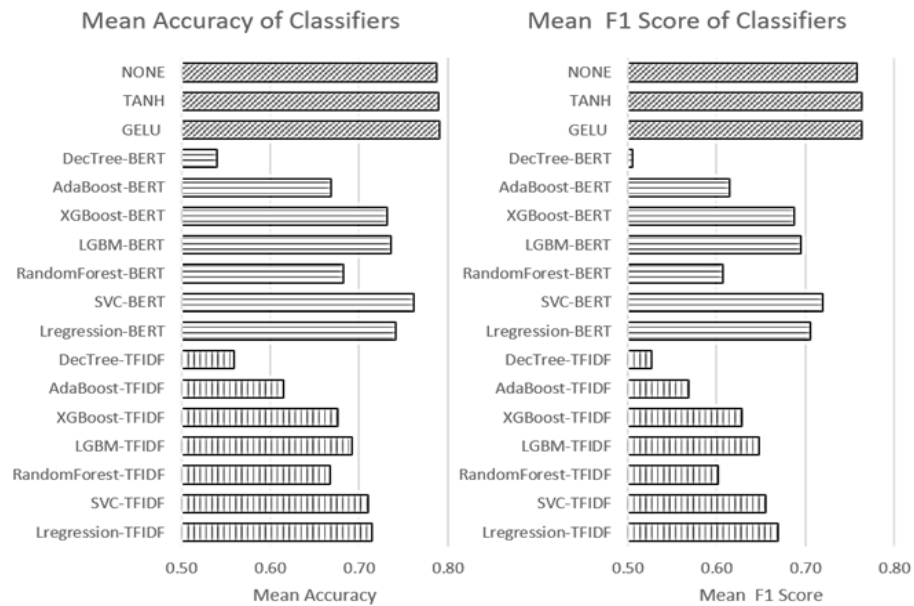


Figure 4. Comparison of mean accuracies and F1-scores for each model

3.2.2. Statistical significance analysis

Figure 5 provides the p-values for pairwise comparisons of the F1-score between all 17 models. All p-values among BERT-based and traditional machine learning models have values less than 0.05, indicating a statistically significant difference. The p-values among BERT-based models are all greater than 0.05, indicating they do not differ significantly. The table also shows that the use of BERT-embedding in traditional machine learning methods mostly differs significantly from using TF-IDF, except for the pair of random forest using BERT embedding and TF-IDF, which has a p-value of 0.3.

V	Models	LR-T	SV-T	RF-T	LG-T	XG-T	AB-T	DT-T	LR-B	SV-B	RF-B	LG-B	XG-B	AB-B	DT-B	GELU	TANH	NONE
TF-IDF	LogR (LR-T)	-	9E-06	4E-08	5E-05	2E-06	2E-09	8E-10	8E-07	1E-09	5E-08	5E-07	3E-04	3E-07	2E-11	3E-08	2E-08	2E-08
	SVC (SV-T)		-	6E-07	3E-02	2E-04	3E-09	6E-10	6E-08	1E-09	9E-07	8E-09	5E-06	4E-06	3E-11	3E-09	2E-09	2E-09
	RF (RF-T)			-	4E-09	7E-05	5E-05	5E-07	2E-09	4E-11	3E-01	6E-09	1E-09	7E-02	9E-08	4E-10	4E-10	1E-09
	LightGBM (LG-T)				-	2E-04	4E-08	7E-09	6E-08	1E-09	4E-06	2E-07	2E-06	1E-04	1E-09	4E-09	7E-09	1E-08
	XGBoost (XG-T)					-	5E-07	1E-07	3E-08	3E-09	8E-03	1E-07	3E-07	2E-02	8E-09	8E-09	9E-09	6E-09
	AdaBoost (AB-T)						-	2E-05	2E-10	4E-11	3E-04	1E-10	9E-11	2E-05	1E-06	2E-11	1E-11	2E-11
	DecTree (DT-T)							-	6E-12	1E-11	6E-07	1E-10	6E-10	2E-08	1E-04	2E-13	2E-12	8E-13
BERT Embedding	LogR (LR-B)								-	8E-05	1E-08	3E-03	4E-03	9E-11	8E-14	2E-07	8E-07	6E-07
	SVC (SV-B)									-	7E-10	5E-07	3E-06	4E-10	6E-13	5E-06	6E-06	1E-05
	RF (RF-B)										-	3E-08	3E-08	3E-01	7E-08	4E-09	7E-09	1E-08
	LightGBM (LG-B)											-	5E-02	6E-09	3E-12	2E-07	1E-07	1E-07
	XGBoost (XG-B)												-	6E-07	2E-10	1E-06	4E-07	7E-07
	AdaBoost (AB-B)													-	1E-11	2E-10	2E-10	1E-10
	DecTree (DT-B)														-	2E-13	4E-13	4E-14
BERT	IndoBERTweet - GELU															-	1E+00	2E-01
	IndoBERTweet - TANH																-	2E-01
	IndoBERTweet - NONE																	-

Figure 5. The p-values for pairwise comparisons between models



### 3.2.3. Word-attribution analysis

This study focuses on sentiment classification using word attribution derived from the integrated gradients method, supplemented by word frequency analysis. The integrated gradients method enhances the interpretability of sentiment predictions by calculating word-level contributions from a baseline to the actual input, particularly in the IndoBERTweet-GELU model, which is the best-performing model in this research. Although other methods like shapley additive explanations (SHAP) [32] and local interpretable model-agnostic explanations (LIME) [33] could provide additional interpretability by analyzing word interactions or building local approximations, they were not included in this study.

#### a. Word frequency clouds

Word frequency clouds visually summarize the most frequently mentioned terms within each sentiment category-negative, neutral, and positive. These clouds help identify dominant themes in public discourse.

- Negative sentiment: As shown in Figure 6(a), the word frequency cloud for negative sentiment highlights terms such as *'kebakaran hutan'* (forest fires) and *'impor beras'* (rice imports). These terms strongly focus on environmental crises and food security, reflecting widespread public concern.
- Neutral sentiment: Figure 6(b) illustrates that neutral sentiment is dominated by words like *'mobil listrik'* (electric cars) and *'stunting'*, which suggest ongoing discussions around technology and public health, typically reported in a factual or neutral tone.
- Positive sentiment: The word frequency cloud for positive sentiment Figure 6(c) emphasizes terms such as *'Indonesia'* and *'kendaraan listrik'* (electric vehicles), indicative of national pride and optimism towards technological progress and environmental sustainability.

It is important to note that these word clouds are based on raw word frequencies and do not account for the specific performance of sentiment analysis models.

#### b. Word attribution clouds

Word attribution clouds, derived from Integrated Gradients, spotlight the words with the highest mean attribution values, significantly impacting sentiment classification.

- Negative sentiment: As depicted in Figure 7(a), words like *'positif covid'* (COVID-positive) and *'kebijakan impor'* (import policy) exhibit high attribution values. These terms are critical in driving negative sentiment, highlighting public concerns over health crises and economic policies.
- Neutral sentiment: Figure 7(b) shows that in neutral sentiment, words such as *'hutan mangrove'* (mangrove forest) and *'obat herbal'* (herbal medicine) have high attribution values. These words are central to neutral discourse, reflecting balanced, context-specific content.
- Positive sentiment: For positive sentiment, Figure 7(c) reveals that words like *'salurkan bantuan'* (distribute aid) and *'jaga ketahanan'* (maintain resilience) hold the highest attribution values. These terms underscore the importance of community support and resilience in positive sentiment expressions.

These attribution clouds are generated using the IndoBERTweet-GELU model, which is the best-performing model in our analysis, thereby providing more accurate and sentiment-specific insights.

#### c. Comparison of word clouds: frequency vs. attribution

Table 2 compares the top four words derived from both frequency and attribution analyses for each sentiment category to understand the differences in word significance based on analysis approaches.

- Negative sentiment: The frequency cloud emphasizes broader societal issues like *'kebakaran hutan'* (forest fires), while the attribution cloud identifies *'positif covid'* (COVID-positive) as a key driver of negative sentiment despite its lower frequency. This result demonstrates the IndoBERTweet-GELU model's ability to discern more impactful words contributing to negative sentiment.
- Neutral sentiment: Common topics like *'mobil listrik'* (electric cars) are prominent in the frequency cloud, whereas *'hutan mangrove'* (mangrove forest) emerges in the attribution cloud, showing its importance in neutral discussions despite its relative infrequency. This result reflects the model's nuanced understanding of neutral sentiment.
- Positive sentiment: While *'Indonesia'* is frequently mentioned in positive sentiment, the attribution cloud emphasizes action-oriented terms like *'salurkan bantuan'* (distribute aid), revealing their deeper emotional impact. This result highlights the refined sentiment analysis capabilities of the IndoBERTweet-GELU model.

The word-attribution analysis using the IndoBERTweet-GELU model shows a marked improvement in the accuracy and depth of sentiment analysis, especially within Indonesian Twitter discussions on biodiversity policy, by overcoming the limitations of traditional frequency-based methods, which often fail to capture contextual depth. Unlike these conventional methods, IndoBERTweet-GELU effectively grasps the contextual relevance of words in Indonesian's unique linguistic and cultural nuances, identifying frequently used terms while accurately attributing sentiment, thus uncovering subtle factors like specific lexical choices that shape emotional responses. These findings are crucial for fields requiring precise sentiment interpretation, such as public opinion on environmental policies, as they enable policymakers to make decisions grounded in

more accurate, actionable insights. This study emphasizes the essential role of advanced models like IndoBERTtweet-GELU in delivering high-quality analytical results and stresses the need for continuous model refinement to meet the complexities of sentiment analysis across diverse linguistic settings.



Figure 6. Word frequency clouds for each sentiment class: (a) negative, (b) neutral, (c) positive

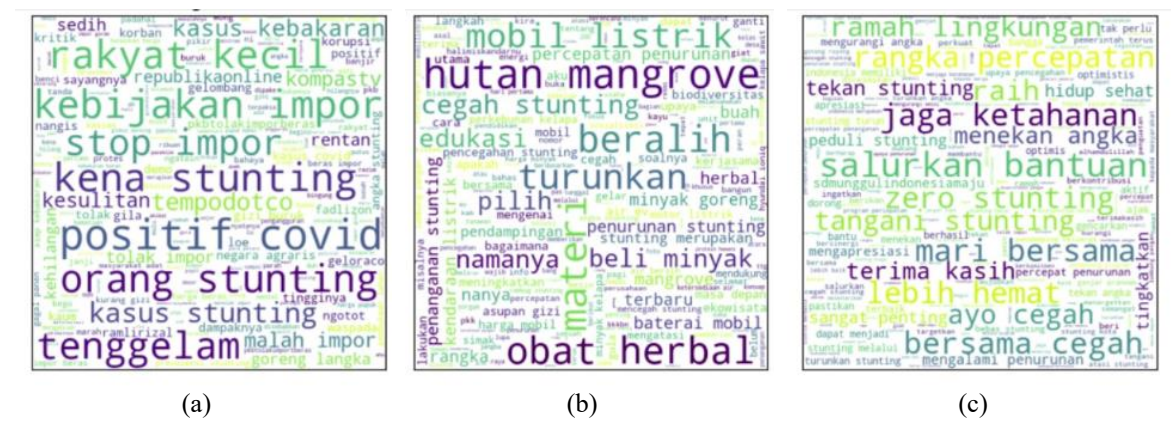


Figure 7. Word Attribution clouds for each sentiment class: (a) negative, (b) neutral, (c) positive

Table 2. Comparison of top words by frequency and attribution		
Sentiment	Top words (frequency)	Top words (attribution)
Negative	'kebakaran hutan' (forest fires), 'impor beras' (rice imports), 'stunting', 'petani' (farmers)	'positif covid' (COVID-positive), 'kebijakan impor' (import policy), 'kena stunting' (affected by stunting), 'rakyat kecil' (common people)
Neutral	'mobil listrik' (electric cars), 'stunting', 'minyak goreng' (cooking oil), 'Indonesia'	'hutan mangrove' (mangrove forest), 'materi' (material), 'obat herbal' (herbal medicine), 'beralih' (switch)
Positive	'Indonesia', 'kendaraan listrik' (electric vehicles), 'mobil listrik' (electric cars), 'angka stunting' (stunting figures)	'salurkan bantuan' (distribute aid), 'jaga ketahanan' (maintain resilience), 'mari bersama' (let's be together), 'bersama cegah' (together prevent)

4. CONCLUSION

This study confirms that IndoBERTtweet with the GELU activation enhances both accuracy and interpretability in sentiment analysis for Indonesian biodiversity policy discourse, establishing it as a valuable tool for public sector evaluations. Applied to 13,435 tweets, IndoBERTtweet-GELU consistently outperformed traditional classifiers, with a mean accuracy of 78.99%, improving by 7.55% over the best TF-IDF model and by 2.83% over other BERT-based models, while Word-Attribution Analysis demonstrated IndoBERTtweet-GELU’s capacity for contextual relevance, providing nuanced sentiment insights within Indonesian Twitter’s unique linguistic context. However, potential biases require attention, as IndoBERTtweet reflects demographic

biases typical of Twitter's younger, urban user base, and its focus on informal language may limit adaptability to formal or regional dialects; data augmentation, model fine-tuning, and a demographic analysis could mitigate these issues. Expanding IndoBERTweet's application to platforms like Facebook and Instagram, developing specialized models for specific biodiversity subdomains, and integrating multimodal data such as images or videos are recommended for future studies to broaden and deepen sentiment analysis. Lessons learned emphasize the importance of meticulous dataset curation to ensure high-quality sentiment labels, as well as clear annotator guidelines to minimize data preparation challenges, as this study underscores IndoBERTweet's potential in supporting sentiment-informed policy decisions, with improvements in platform scope and multimodal capabilities promising further advancements.

## FUNDING INFORMATION

The authors declare that no funding was received to support the conduct of this study.

## AUTHOR CONTRIBUTIONS STATEMENT

The authors' individual contributions to this work are categorized according to the Contributor Roles Taxonomy (CRediT) as follows:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohammad Teduh Uliniansyah	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Asril Jarin	✓	✓					✓	✓	✓	✓			✓	
Agung Santosa	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		
Gunarso				✓		✓	✓			✓				

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## DATA AVAILABILITY

The data supporting the findings of this study are openly available in the Mendeley Data repository at <https://data.mendeley.com/datasets/xtk9wsxjrr/4> [11].

## REFERENCES




- [1] K. V. Rintelen, E. Arida, and C. Häuser, "A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries," in *Research Ideas and Outcomes*, Sep. 2017, doi: 10.3897/rio.3.e20860.
- [2] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: a pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 10660–10668, 2021, doi: 10.18653/v1/2021.emnlp-main.833.
- [3] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.
- [4] A. D. Latief, A. Jarin, M. T. Uliniansyah, E. Nurfadhilah, and D. I. N. Afra, "A proven sentiment annotation guideline for Indonesian Twitter data," in *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, pp. 31–36, Oct. 2023, doi: 10.1109/IC3INA60834.2023.10285807.
- [5] A. Santosa, A. Jarin, T. Sampurno, and S. Pebiana, "Top layer selection in pretrained models for sentiment analysis on biodiversity Tweets in Bahasa Indonesia," in *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, pp. 149–154, Oct. 2023, doi: 10.1109/IC3INA60834.2023.10285786.
- [6] A. Bello, S.-C. Ng, and M.-F. Leung, "A BERT framework to sentiment analysis of Tweets," *Sensors*, vol. 23, no. 1, Jan. 2023, doi: 10.3390/s23010506.
- [7] S. Singh, K. Kumar, and B. Kumar, "Sentiment analysis of Twitter data using TF-IDF and machine learning techniques," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, pp. 252–255, May 2022, doi: 10.1109/COM-IT-CON54601.2022.9850477.
- [8] A. Samir, S. M. Elkaffas, and M. M. Madbouly, "Twitter sentiment analysis using BERT," in *2021 31st International Conference on Computer Theory and Applications (ICCTA)*, IEEE, pp. 182–186, Dec. 2021, doi: 10.1109/ICCTA54562.2021.9916614.

*Modeling sentiment analysis of Indonesian biodiversity policy Tweets ... (Mohammad Teduh Uliniansyah)*

- [9] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on Twitter using streaming API," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, IEEE, pp. 915–919, Jan. 2017, doi: 10.1109/IACC.2017.0186.
- [10] M. T. Uliniansyah *et al.*, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data in Brief*, vol. 52, Feb. 2024, doi: 10.1016/j.dib.2023.109890.
- [11] M. T. Uliniansyah *et al.*, "Indonesian biodiversity-related tweets including health, food security, and environmental management issues for sentiment analysis," *Mendeley Data*, ver. 4, doi: 10.17632/xtk9wsxjir.4
- [12] S. Pebiana *et al.*, "Experimentation of various preprocessing pipelines for sentiment analysis on Twitter data about new Indonesia's capital city using SVM And CNN," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, pp. 1–6, Nov. 2022, doi: 10.1109/O-COCOSDA202257103.2022.9997982.
- [13] Sastrawi, "Sastrawi: Python-based Text Stemmer for Bahasa Indonesia," *Python Package Index (PyPI)*. Accessed: Oct. 18, 2024. [Online]. Available: <https://pypi.org/project/Sastrawi/>
- [14] C. E. McCulloch, "Generalized linear models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1320–1324, Dec. 2000, doi: 10.1080/01621459.2000.10474340.
- [15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [16] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 1–9.
- [17] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995, pp. 278–282 vol.1, doi: 10.1109/ICDAR.1995.598994.
- [18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA: ACM, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [19] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 37–52, doi: 10.1007/978-3-642-41136-6\_5.
- [20] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [22] C. Schaffer, "Selecting a classification method by cross-validation," *Machine Learning*, vol. 13, no. 1, pp. 135–143, Oct. 1993, doi: 10.1007/BF00993106.
- [23] H. Hsu and P. A. Lachenbruch, "Paired *t* Test," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014, doi: 10.1002/9781118445112.stat05929.
- [24] J. Lei, "Cross-validation with confidence," *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1978–1997, Oct. 2020, doi: 10.1080/01621459.2019.1672556.
- [25] D. Cox, "Statistical significance tests," *British Journal of Clinical Pharmacology*, vol. 14, no. 3, pp. 325–331, Sep. 1982, doi: 10.1111/j.1365-2125.1982.tb01987.x.
- [26] K. Chu, "An introduction to statistics, significance testing and the P value," *Emergency Medicine*, vol. 11, no. 1, pp. 28–34, Mar. 1999, doi: 10.1046/j.1442-2026.1999.00316.x.
- [27] F. Emmert-Streib and M. Dehmer, "Understanding statistical hypothesis testing: The logic of statistical inference," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 945–961, Aug. 2019, doi: 10.3390/make1030054.
- [28] V. Bartíčka, O. Pražák, M. Konopík, and J. Sido, "Evaluating attribution methods for explainable NLP with transformers," *Proceedings, Text, Speech, and Dialogue: 25th International Conference*, TSD 2022, Brno, Czech Republic, 2022, pp. 3–15, doi: 10.1007/978-3-031-16270-1\_1.
- [29] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv-Computer Science*, pp. 1–11, Mar. 2017.
- [30] G. Khanvilkar and P. Deepali Vora, "Sentiment analysis for product recommendation using random forest," *International Journal of Engineering & Technology*, vol. 7, no. 3.3, pp. 87–89, Jun. 2018, doi: 10.14419/ijet.v7i3.3.14492.
- [31] S. Kiritchenko and S. M. Mohammad, "Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 811–817, 2016, doi: 10.18653/v1/N16-1095.
- [32] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv-Computer Science*, pp. 1–10, May 2017.
- [33] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA: ACM, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.




## BIOGRAPHIES OF AUTHORS






**Mohammad Teduh Uliniansyah**    received a B.Eng. degree from Shibaura Institute of Technology, Japan, in 1991, a Master of Computer Science degree from Oklahoma State University, USA, in 1998, and a Ph.D. from Keio University, Japan, in 2007. He holds the associate researcher position at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 1992, he has been actively engaged in research within artificial intelligence, particularly on speech and natural language processing. His research encompasses speech recognition and analysis, sentiment analysis, and natural language processing. He can be contacted at email: [mtd001@brin.go.id](mailto:mtd001@brin.go.id).








**Asril Jarin**    holds a Doctor of Electrical Engineering from the University of Indonesia in 2017. He also received his Dipl.Ing from GSO Fachhochschule Nuernberg and M.Sc. from Hochschule Darmstadt, Germany, in 2001 and 2003 respectively. Presently, he serves as an associate researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in the domain of artificial intelligence, specifically focusing on speech and natural language processing. His research includes speech recognition, speech analytics, sentiment analysis, and natural language processing. Over time, he has contributed with numerous publications in international conferences and journals. He can be contacted at email: [asri003@brin.go.id](mailto:asri003@brin.go.id).



**Agung Santosa**    received the BSEE from University of Toledo, USA, in 1992 and Master of Computer Science from University of Indonesia, Indonesia, in 2015. He is an associate researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in the domain of artificial intelligence, specifically focusing on speech and natural language processing. His research includes speech recognition, speech analytics, sentiment analysis, and natural language processing. He can be contacted at email: [agun006@brin.go.id](mailto:agun006@brin.go.id).



**Gunarso**    graduated in electrical engineering from Gadjah Mada University (UGM) Indonesia in 1988. He is an associate researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in the domain of artificial intelligence, specifically focusing on speech and natural language processing. His research includes speech recognition, sentiment analysis, and natural language processing. He can be contacted at email: [gunarso@brin.go.id](mailto:gunarso@brin.go.id).