PTI JOURNAL
LPPM Universitas Putra Indonesia YPTK Padang
Lubuk Begalung Highway, Padang, Sumatera Barat, Indonesia
Volume: 13, Issue: 2, Page: 78 - 82, 30/10/2026, e-ISSN: 2685-3914

Classification of Indonesia Texts on X Related to Environmental Issues

Ridea V.P. Siwabessy[1], Muhammad Okky Ibrohim[1,2*], Indra Budi[1], Naufal Zafran Fadil[1]
[1] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia
[2] Institute of Social Studies, Faculty of Social Sciences, University of Tartu, Tartu, 51003, Estonia
ridea.v@ui.ac.id, okkyibrohim@cs.ui.ac.id, indra@cs.ui.ac.id, naufal.zafran@ui.ac.id

Abstract

Environmental issues have become a global concern, including Indonesia. X is one of the social media that is often used to discuss environmental issues. Text Classification is needed to identify and categorize tweets that are relevant to environmental issues, thus facilitating the further analysis process. This research aims to develop a text classification system related to environmental issues written in Indonesian. The data used in this research are tweets from X related to environmental issues. The pre-processing techniques performed are case folding, text normalization, stopword removal, and stemming before classification. The classification algorithms used are Logistic Regression (LR), Random Forest (RF), Long Short-Term Memory (LSTM) and IndoBERT pre-trained model. Based on the research results, the IndoBERT-base-p2 model provides the best performance in the classification of Indonesian texts related to environmental issues with an F1-Score of 87,478%.

Keywords: environmental issues, IndoBERT pre-trained model, natural language processing, text classification, social media X

Jurnal PTI is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

1. Introduction

Environmental issues are currently a global concern, and Indonesia is no exception. There are many important topics or issues related to the environment itself. Through the UN IPCC report in 2018, it was explained that one of the environmental problems is climate change which will occur within 12 years [1]. In addition, human activities have caused many environmental problems that are exacerbating and threatening such as deforestation, ecosystem damage, pollution, water and soil pollution [2]. The United Nations has sought various ways to address these issues, one of which is by setting the Sustainable Development Goals (SDGs)[1] in 2015 to achieve a more just, prosperous and sustainable world by 2030. One of the main focuses of the SDGs is to maintain environmental sustainability, which is reflected in several goals related to environmental issues.

X[2] is one of the social media that is often utilized by the public to share information with users. X's straightforward features make it easy for users to express their opinions and short reports related to a particular issue or topic. In addition, another advantage of X social media is the ability to spread tweets very quickly because there is no filtering, so many topics or issues that are trending in the real world can be discussed immediately on X social media. One of the topics or issues discussed is environmental issues. In relation to, social media X is often used to spread information about environmental issues not only to environmental activists but also to the wider community [3].

As the volume of information data on X grows, so does the need to obtain information more effectively and efficiently. In this context, text classification becomes a very important technique in facing these challenges [4]. In particular, text classification focuses on categorizing and organizing text data for management purposes and simplifying the analysis process [5]. In this case, text classification can be used to identify and categorize tweets that are relevant and irrelevant to environmental issues.

In the context of NLP, various methods of both machine learning and deep learning have been carried out to perform text classification. [6] performed sentiment classification form Amazon product review data using the Logistic Regression (LR) algorithm. The research used several feature extraction methods. [7] also performed Indonesian text classification on e-learning user feedback data at a university. The study used six classification methods including Logistic Regression

[1] https://www.un.org/sustainabledevelopment/

[2] https://x.com/home

Accepted: 20-08-2026 | Revision: 22-09-2026 | Publication: 30-10-2026 | doi: 10.35134 1

(LR), Decision Tree (DT), Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (kNN), Linear SVC and Random Forest (RF). The results show that in general, Logistic Regression has better performance compared to other classification algorithms.

[8] utilized deep learning methods to conduct news text classification research using the Long Short-Term Memory (LSTM) algorithm. In another study conducted by [9] showed that LSTM has the highest accuracy with 92% compared to the Support Vector Machine (SVM) and RF algorithms for the task of classifying Javanese text expressions. In addition, Bidirectional Encoder Representation from Transformers (BERT) [10], [11], [12] has been used to design text classification models although not specifically designed for the task. Vectorization using BERT is context-dependent resulting in better word vector representation [13]. In terms of the categorization, the BERT model is superior and has better performance compared to other methods [14].

In its development, the researchers then designed a pre-trained IndoBERT model [15] to process various natural language processing tasks in Indonesian including text classification [16].

In this research, the text classification of X tweets is caried out utilizing machine learning methods (LR and RF), deep learning methods (LSTM and IndoBERT pre-trained models). This research aims to analyse and compare the classification results obtained from the various methods applied using the best F_1-Score value

2. Methods

This section details the text classification process related to environmental issues. The flow of our research experiment can be seen in Figure 1

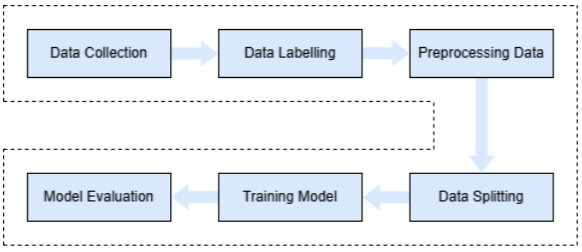


Figure 1. Research Experiment Flow

2.1 Data Collection

The first step in the tweet classification process is data collection. In this study, we used X data in the form of Indonesian tweets. The data collected is text data. The data is collected using the Twitter API which is implemented using the Tweepy Library. The data is collected based on queries or keywords which are words, phrases, or hashtags related to environmental issues. The keywords were adapted into Indonesian from several keywords [17].

2.2 Data Labelling

After the data collection process, the next step is the data labelling process. The data labelling process is carried out by self-annotating 600 tweets consisting of 300 tweets that discuss environmental topics and 300 tweets that do not discuss environmental topics.

Table 1. Tweet reference criteria for data labelling

Discuss Enviromental Issues	Not Discuss Enviromental Issues
Normal tweet discusses environmental issues;	Normal tweet that does not discuss environmental issues;
Tweets from official accounts discussing public service advertisements related to environmental issues;	Tweets from official accounts that do not discuss public service announcements related to environmental issues;
Tweets from official accounts that discuss news related to environmental issues.	Tweets form official accounts that do not discuss news related to environmental issues;
	Tweets that discuss environmental management products;
	Tweets that do not discuss environmental management products;
	Spammy tweets;
	Tweets that contain fraud.

2.3 Preprocessing Data

In this process, data preprocessing is carried out, namely case folding, text normalization, stopword removal, and stemming. This is done to prepare and clean the data before the next process for training the model.

2.4 Data Splitting

After preprocessing, data division is then carried out with a composition of 80:20 where 80% of the dataset is used as initial train data and the 20% is used as test data. Furthermore, 80% of the train data is divided again with ratio of 80:20, so that 80% of the initial train data is used as train data and the 20% is used as validation data.

2.5 Training Model

This research uses several classification models namely LR, RF, LSTM and IndoBERT pre-trained models.

2.5.1 Logistic Regression (LR)

Logistic Regression models the relationship between features and the probability of belonging to a particular

class using a logistic function [18]. Before applying LR, each text is first converted into numerical features or feature vectors. Feature vectors are represented based on the frequency or frequency weighted values of the words or phrases in the text. Then LR models the prediction class probabilities into feature vectors which are inputs to the LR model. A Logistic function is applied to map real numbers into a range of 0 to 1 to represent probabilities. The trained model can be used for text classification into specific classes, by first determining a threshold value. Subsequently, the predicted probabilities will be compared against the threshold value to determine the class of the text.

2.5.2 Random Forest (RF)

Random Forest is an ensemble learning method, which is a method that combines several classification algorithms (either different algorithms or the same algorithm but done repeatedly). According to [19] the classification process with RFDT is carried out by combining several decision trees which then the classification results are determined based on the majority voting of the decision trees formed.

2.5.3 Long Short-Term Memory (LSTM)

LSTM is one of the improved methods of Recurrent Neural Network (RNN). For the first time, Hochreiter and Schmidhuber developed the LSTM algorithm [8]. The LSTM algorithm was designed to overcome the weakness of RNN which cannot predict words stored in long-term memory. LSTM is designed to overcome the weaknesses, but still maintain the advantages of RNN. This is done because RNN is able to provide accurate predictions of the information. With its set of gates, the LSTM is able to store important relevant information and find patterns for a long period of time. This then allows the LSTM to be used to process, predict, and classify information based on time series data. The gates owned by LSTM consist of input gate, forget gate and output gate. The input gate updates the state and decides which information is important and unimportant. Forget gate decides which information is kept and which is irrelevant so that it can be discarded. The output gate is the final stage of the circuit which serves to extract useful information to be presented as output values. The sigmoid and tanh functions play a very important role in the input and output gates in identifying relevant information as well as being a filter.

2.5.4 Indonesia Bidirectional Encoder Representation Transformer (IndoBERT)

IndoBERT is pre-trained BERT model [11] adapted to understand and process Indonesian text data [15]. IndoBERT has been trained on a large dataset called Indo4B. The dataset was collected from three sources, namely Wikipedia, news articles, and Indonesian web corpus. The IndoBERT training process was conducted in 2 phases. The first training phase, IndoBERT used a maximum sequence length of 128 and the second training phase used a maximum sequence length of 512. The second phase of training refines the first phase allowing IndoBERT to learn even more.

2.6 Model Evaluation

The classification results are then evaluated using the accuracy, precision, recall and F1-Score evaluation metrics. The evaluation can be seen in Figure 2 and equations (1)-(4). The most optimal model will be selected based on the highest F1-Score value on the test data.

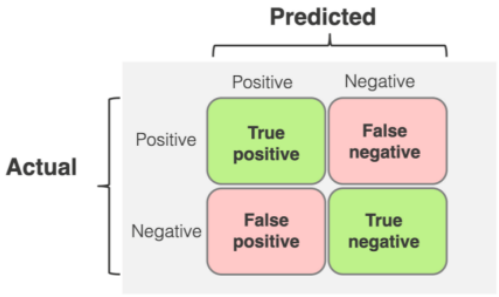


Figure 2. Confusion Matrix [20]

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 - Score = 2 \times \frac{Recall \times Precission}{Recall+Precission} \quad (4)$$

- True Positive (TP) : A condition if the prediction result is positive then the actual value is also positive so the classification is correct.
- True Negative (TN) : A condition where the predicted label is negative and the actual label is negative so the classification is correct.
- False Positive (FP) : A condition where the prediction result gives a positive label even though the actual label is negative so the classification is wrong.
- False Negative (FN) : A condition where the predicted label is negative and the actual label is positive, resulting in an incorrect classification.

3. Results and Discussions

This section discusses the results of our study along with error analysis of the best model results.

3.1 Classification Model Building Results

Tweet X classification task is performed using the four algorithms mentioned in Chapter 2, namely LR, RF, LSTM and IndoBERT. The classification model that has been built is then evaluated using evaluation metrics. The evaluation results can be seen in Table 2

Table 2. Performance comparison of the algorithms

Algorithm Name	Accuracy	Precision	Recall	F1-Score
LR+Unigram	81,667%	83,111%	81,119%	81,245%
RF+Unigram	81,667%	84,316%	80,952%	81,029%
LSTM+Word 2Vec	80%	81%	80%	80%
IndoBERT-base-p2	**87,5%**	**87,51%**	**87,458%**	**87,478%**

From Table 2, it can be seen that the performance of the IndoBERT-base-p2 model is superior to other algorithms based on the evaluation results conducted on test data in classifying X tweets related to environmental issues.

3.2 Classification Model Error Analysis

Error analysis in this study was carried out by evaluating the confusion matrix [20] generated from the classification process on each label. The confusion matrix in this study can be seen in Figure 3.

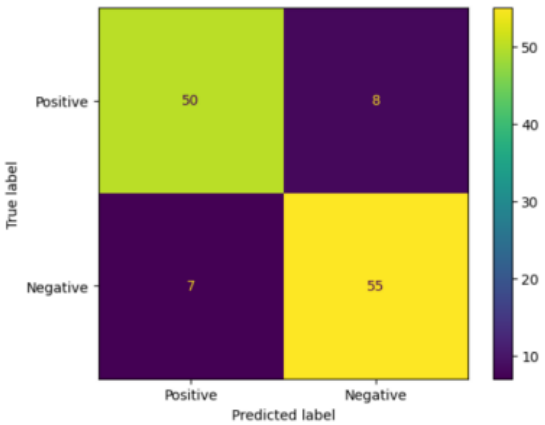


Figure 3. Confusion matrix of IndoBERT-base-p2 model

Table 3 is an example of the prediction error made by the model

Table 3. Examples of Misclassified Sentiment Analysis Results

No	Text	Actual Label	Prediction Label
1	Wah kacau banget ya iklim dunia skrg, efek global warming sngat terasa, cuaca ga menentu dan ga bersahabat, semoga para pecinta lingkungan hidup makin semangat aja deh jaga bumi kita Jaga keaehatan teh '	Positive	Negative
2	KST Papua Rusak Kedamaian di tanah Papua #KSTPapua #TerorKSTPapua #PapuaIndonesia #DamaiaPapua #TolakKSTPapua https://t.co/orFI5aN1jz'	Negative	Positive

Based on Table 3. It was found that the prediction error in text 1 occurred because the model failed to capture the full context of environmental issues. This is also because the sentence contains general and social messages as well as motivations that confuse the model to predict correctly. The prediction error in text 2 is likely to occur due to the use of hashtags that confuse the model in performing a precise classification. Although the sentence clearly discusses security and social conflict in Papua, the model may associate it with several keywords related to environmental issues. This may happen because the data train is not fully representative.

4. Conclusions

X is one of the platforms used to share messages and access information quickly related to environmental issues. Of the four algorithms used to build a text classification model related to environmental issues, it was found that the IndoBERT-base-p2 model was superior to the other three algorithms. This can be seen from the F1-score value of 87,478%. Prediction errors during the tweet classification process occur because the model fails to capture the full context of environmental issues. This is due to the limited number of data train that are not representative, causing prediction errors. For future work, we suggest several improvements. First, the model can be trained or fine-tuned on more, diverse and representative text data with social language styles that discuss environmental issues. Second, specifically for Twitter data, it can be trained using a specific variant of the IndoBERTweet model to better complete tasks with tweet data sets.

5. References

[1] IPCC, "Global warming of 1.5° C: an IPCC special report on the impacts of global warming of 1.5° C above pre-industrial levels and related global greenhouse gas emission pathways," 2018.

[2] N. Hadiyati and Cindo, "KONTEKSTUALISASI PENCEMARAN EKOSISTEM LAUT DALAM MENCAPAI SDGS: SUATU KAJIAN HUKUM LINGKUNGAN DI INDONESIA," *Nusantara: J. Ilmu Pengetah. Sos.*, vol. 8, 2021, doi: 10.31604/jips.v8i3.2021.300-313.

[3] N. Zhang and M. M. Skoric, "Making the news: environmental NGOs and their media visibility in China," *Chin. J. Commun.*, pp. 395–413, 2018.

[4] Q. Li et al., "A Survey on Text Classification: From Traditional to Deep Learning," *Assoc. Comput. Mach.*, Apr. 2022, doi: 10.1145/3495162.

[5] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications," *Comput. Sci. Rev.*, Nov. 2024.

[6] P. S. Reddy, D. R. Sri, S. R. Cheerka, and S. Shaik, "Sentimental Analysis using Logistic Regression," *Int. J. Eng. Res. Appl.*, vol. 11, pp. 36–40, Jul. 2021.

[7] A. Rusli, A. Suryadibrata, S. B. Nusantara, and J. C. Young, "A Comparison of Traditional Machine Learning Approaches for Supervised Feedback Classification in Bahasa Indonesia," *Int. J. New Media Technol.*, vol. VII, no. 1, pp. 28–32, Jun. 2020.

[8] I. Triyadi, B. Prasetiyo, and T. L. Nikmah, "News text classification using long-term short memory (LSTM) algorithm," *J. Soft Comput. Explor.*, 2023.

[9] O. V. Putra, A. Musthafa, and K. P. Wibowo, "Klasifikasi Ekspresi Teks Berbahasa Jawa Menggunakan Algoritma Long Short Term Memory," *Komputika*, vol. 10, no. 2, pp. 137–143, Aug. 2021, doi: 10.34010/komputika.v11i1.4616.

[10] M. A. Shah, M. J. Iqbal, N. Noreen, and I. Ahmed, "An Automated Text Document Classification Framework using BERT," *IJACSA*, vol. 14, no. 3, 2023.

[11] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT 2019*, Jun. 2019, pp. 4171–4186.

[12] E. Demir and M. Bilgin, "Sentiment Analysis from Turkish News Texts with BERT-Based Language Models," in *2023 8th Int. Conf. Comput. Sci. Eng. (UBMK)*, IEEE, 2023, pp. 147–150, doi: 10.1109/UBMK59864.2023.10286719.

[13] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 305–310, 2019.

[14] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *J. Comput. Cogn. Eng.*, vol. 2, no. 4, pp. 352–356, Nov. 2023, doi: 10.47852/bonviewJCCE3202838.

[15] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proc. ACL-IJCNLP*, Dec. 2020, pp. 843–857.

[16] B. Juarto and Yulianto, "Indonesian News Classification Using IndoBert," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 454–460, 2023.

[17] M. O. Ibrohim, C. Bosco, and V. Basile, "Sentiment Analysis for the Natural Environment: A Systematic Review," *ACM Comput. Surv.*, vol. 56, no. 4, Apr. 2023, doi: 10.1145/3604605.

[18] M. Maalouf, "Logistic regression in data analysis: An overview," *Inderscience Publishers*, 2011, doi: 10.1504/IJDATS.2011.041335.

[19] R. Khan, N. Ahmad, J. Ali, and I. Maqsood, "Random Forests and Decision Trees," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, 2012.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[21] A. Nursandi and A. S. Girsang, "Waste Pollution Classification in Indonesian Language using DistilBERT," *Gema Wiralodra*, vol. 15, no. 1, 2024.

[22] I. Octovianto, M. O. Ibrohim, and I. Budi, "Indonesian Sentiment Analysis in Natural Environment Topics," *Indones. J. Electr. Eng. Comput. Sci.*, 2025.

[23] Fontanella et al., "Sentiment Analysis of Indonesian Government Policy in the Environmental Sector," 2021.

[24] M. Rochyadi-Reetz and J. Wolling, "Environmental Communication Publications in Indonesia's Leading Journals: A Systematic Review," 2026.

[25] Lintang et al., "Early Detection of Environmental Issues from Social Media using IndoBERT and LDA," 2025.