

# Indonesian sentiment analysis in natural environment topics

Christofer Octovianto<sup>1</sup>, Muhammad Okky Ibrohim<sup>1,2</sup>, Indra Budi<sup>1</sup>

<sup>1</sup>Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

<sup>2</sup>Dipartimento di Informatica, Università Degli Studi di Torino, Turin, Italy

---

## Article Info

### Article history:

Received Apr 6, 2024

Revised Nov 12, 2024

Accepted Nov 24, 2024

---

### Keywords:

Data-driven analyses

IndoBERT model

Instagram

Natural environment topic

Sentiment analysis

---

## ABSTRACT

Indonesia is one of the countries that is rich in biodiversity and has a high population growth. This condition can cause Indonesia to have problems related to the natural environment that are more complex than other countries. Hence, this has created a lot of discussions regarding natural environmental issues in Indonesia on social media platforms. In this case, stakeholders like the government in general can utilize sentiment analysis (SA) to comprehend the public's views to allow them to better fit the public's expectations when formulating a particular policy that related to the environmental sustainability (ES) issues. This paper built the first open dataset of Indonesian SA dataset in ES topics collected from Instagram. As the benchmark of our dataset, we used IndoBERT model variant for constructing the model and the experiment result shows that model based on IndoBERT-large-p2 obtained the best performance with 72.44% of F1-score.

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



---

## Corresponding Author:

Muhammad Okky Ibrohim

Faculty of Computer Science, Universitas Indonesia

Depok, Indonesia

Email: okkyibrohim@cs.ui.ac.id, muhammadokky.ibrohim@unito.it

---

## 1. INTRODUCTION

Environmental sustainability (ES) is one topic that has received more attention in this era. This takes place mainly because its influence all aspects of life especially in health and life quality. The issues in this domain must be discussed and solved together collectively by the governments and the public. Governments have more crucial role since they can propose and control the ES's policies and regulations. The Public also has an essential role in terms of supervising and evaluating the government's policies and actions, and reporting to their government if there are ES issues around them. In the current social media era, citizens often discuss and report an ES issue on their social media. Hence, the relationship between people and the environment through sentiment analysis (SA) may be crucial to be explored [1].

SA has been broadly applied to several topics SA on patient feedback [2], [3], public facility reviews [4], [5], and product reviews [6], [7], to gauge how people perceive services, products, or other specific discourse topics. Concerning environmental issues, SA can help us comprehend the public's views of the condition of the environment and the government environmental policies that are happening. This can allow governments to better fit the public's expectations when they decide on what needs to be fixed for environmental issues in their representative.

In recent years, some research has been done to explore SA applied to ES topics from various perspectives with various approaches. Ibrohim *et al.* [1], a systematic review of SA for ES topics is conducted to explore what tasks, techniques, and benchmarks have been used in this growing research area. Dahbi *et al.*

[8] have researched SA in the Moroccan language related to the smart city in Morocco. Zhang *et al.* [9] conducted sentiment classification research related to public opinions about food safety in China. They used the IFoodCloud application to gather public opinions and used machine learning and deep learning algorithms to classify the data.

Indonesia is the country with the second most biodiverse country [10] and the fourth largest population [11]. This condition can cause Indonesia to have problems related to the ES that are more complex than other countries because there will be a lot of interaction between humans and the natural environment components. This situation has also created a lot of discussions regarding natural environmental issues in Indonesia in newspapers, blogs, and social media. On social media, users engage in online interaction which leads to the generation of a substantial volume of user-generated information. In addition, social media users come from all over the world with different interests and mindsets [12]. The development of this vast user-generated content can improve the management of existing ES monitoring systems. Based on survey conducted by [13], Instagram is the most used social media for discussing and campaigning on natural environment issues (especially climate change) among users with an age range 20-29 years old, which is the age range of users with the most social media exposure in Indonesia. Therefore, we can get a wide variety of information, specifically in this study for collecting data and analyze the sentiment with regards to the ES issues in Indonesia.

To the best of our knowledge, research for SA applied in ES related topics in Indonesian social media is still very rare. Michael and Utama [14] used Naive Bayes (NB) for conducting SA in waste management cases by utilizing Indonesian tweets. Indra *et al.* [15] used the k-nearest neighbor (KNN) algorithm to comprehend the public view of social and political orientation in Pekanbaru City by utilizing Indonesian tweets. One of the social orientations used is environmental issues in Pekanbaru City. Sugiharti and Fauziah [16] used KNN and maximum entropy methods and Indonesian tweets to gauge public views on menstrual cups to reduce menstrual waste. Fontanella *et al.* [17] used NLTK to analyze the sentiment of the public in Indonesian tweets concerning the implementation of environmental protection and management of coal waste. For research in Indonesian SA about ES topics, unfortunately, we found that all previous works do not share their dataset with the public so other researchers cannot reproduce and enhance their research.

In the terms of classifier used, we found that most of the algorithms previously mentioned rely on heavy feature engineering and pose challenges in terms of optimization. SA, especially using traditional machine learning algorithms, fails to capture the context or the meaning in entire sentences of tweet data due to a limited understanding of semantics and word sequence as they only see certain patterns. This situation requires additional preprocessing or feature engineering. Therefore, this makes analysis more difficult for an algorithm to accurately capture the intended meaning and context of a sentence [18], [19].

Nowadays, most of the state-of-the-art of SA tasks have been achieved by using a model based on bidirectional encoder representations from transformers (BERT) pre-trained model [20]–[23]. In this paper, we focus on the implementation of Indonesian BERT (IndoBERT) [24]. In summary, the main contributions of our paper are as follows:

- Building the first open dataset and model for ES vs non-ES topic classification that is useful for ES data filtering before SA annotation;
- Building the first open dataset for Indonesian Instagram SA in SE topics where each data is annotated by three native annotators;
- Providing benchmark experiment for our dataset including error analysis to help the community in conducting a good system for ES monitoring.

We release our code and dataset in [25]. The rest of this paper is organized as follows. Section 2 discusses related works, section 3 discusses the step for building the dataset and explaining how our experiments are conducted, section 4 presents the experiment results and analysis. Finally, section 5 discusses the conclusions and future works suggestions of this paper.

## 2. RELATED WORKS

Studies related to SA in ES topics have increased in the past few years. Dahbi *et al.* [8] conducted SA using Twitter and Facebook data in the Moroccan language regarding the smart cities in Morocco. The characteristics of smart city in the dataset include pollution, congestion, security risks, and urban resources. They used support vector machine (SVM), NB, KNN, and decision tree (DT) algorithms to build the SA model. Binary weighting for feature extraction is used in this research. The highest accuracy value was achieved by

using the SVM algorithm with an accuracy of 94.1%. Zhang *et al.* [9], conducted SA about public opinions of food safety in China such as views on food adulteration, food-borne diseases, agricultural pollution, irregular food distribution, and problems with food production. The data used was taken from IFoodCloud, a platform built for real-time SA of public opinion on food safety in China. They collect about 3,100 data for experiments and built a model using the lexicon-based algorithm, SVM, multilayer perceptron (MLP), and long short-term memory (LSTM). The best model is achieved by the LSTM algorithm with 97.37% of F1-score.

In Indonesia, there also are studies involving SA on ES topics. Michael and Utama [14], conducted SA concerning waste management cases in Indonesia with the help of the KNN algorithm. The cases were collected from Indonesia tweets with the keywords of *sampah* (waste), *polusi* (pollution), *lingkungan* (environment), and *manajemen* (management). The KNN model achieved an accuracy of 65.02% for this task. Indra *et al.* [15] utilized the KNN algorithm to comprehend the social orientation of the public in Pekanbaru city. One of the public orientations discussed was environmental issues. The data were collected from Indonesian tweets. Environmental issues were the social orientation component that received the most negative sentiment (56% of the environmental data collected). Based on the discourses, “*sampah*” (garbage, waste, and trash) was the most posted and discussed word on the data along with “*banjir*” (flood) and “*polusi udara*” (air pollution). Sugiharti and Fauziah [16] built an SA model based on maximum entropy and KNN algorithms to comprehend public views on menstrual waste and the utilization of menstrual cups. They used 1,108 Indonesian tweets for building the model. The best model was achieved based on maximum entropy with an accuracy of 84.6%. Fontanella *et al.* [17] used NLTK to analyze the sentiment of the public concerning the implementation of environmental protection and management of coal waste. They collected data from Indonesian tweets with a total of 236 tweets. The results showed that the public opinions tended to be negative sentiments with an accuracy rate is 77.40%.

From our literature review process and also stated in a survey by [1] findings, there is one big issue for the NLP communities in building an SA model in ES topics, i.e. the existence of an open dataset. Especially for the Indonesian dataset, all previous research only works in small datasets in a specific ES topic where they do not share the dataset with the public. Therefore, as a country that potentially has bigger ES issues than other countries as mentioned in section 1, there is a need to build an open dataset that widely covers ES topics issue in this language so that the community can help the Indonesian government in conducting a good ES monitoring system.

In 2023, Bosco *et al.* [26] created a complete annotation guideline for collecting and annotating structured sentiment analysis (SSA) data in ES topics. They used 12 keywords to collect Italian Twitter data and 120 keywords to collect 120 English Twitter data. The dataset collected covers all 10 ES topics defined by [1] namely sustainability, environment, green, organism, pollution, waste, food, energy, carbon, and climate change. For the annotation process, they annotate their dataset in two schemas. For the Italian dataset, they perform complete SSA annotation where a tweet is annotated in span and relation tuple label (holder, target, sentiment term, sentiment polarity, and topic). Meanwhile, for the English dataset, they annotate the sentiment term and its polarity in the form of relation tuple (sentiment term and sentiment polarity) and plan to continue to get full annotation scheme as they did in the Italian data in the second stage. In this case, both Italian and English datasets are annotated by three annotators for each data, where they use a native expert annotator for Italian data and use native crowdsourcing annotator from Prolific [27] for English data. For the evaluation, they use Fleiss' [28] and Cohen's  $\kappa$  [29] score and show that both datasets they built are valid for the experiment. Unfortunately, their work is only focused on dataset building without experimenting. They also do not discuss how to aggregate the final label as they annotate at span-level using more than one annotator. For span-level aggregation, Barbieri *et al.* [30] annotated the SA dataset with the same schema did by [26] in their English dataset. In this case, Barbieri *et al.* [30] convert the span-level annotation to the document-level annotation via majority voting between annotated sentiment terms.

As mentioned in section 1 and also stated in survey findings conducted by [1], most of the existing SA in ES topics modeling still rely on lexicon-based libraries and classic machine learning algorithms. Meanwhile, the current state-of-the-art in SA modeling relies on a BERT-based pre-trained model. For SA in Indonesian data, Wilie *et al.* [24] shows that the BERT-based model outperforms all classic machine learning and classic deep learning models in the SA task in the general topic. In this case, they train various BERT-based models named IndoBERT and share the pre-trained model with the public so that the NLP community can fine-tune it for a specific downstream task. Therefore, it is interesting to use this IndoBERT-based model for Indonesian SA in ES topics modeling as a benchmark.

### 3. METHOD

This section details the process of applying SA for the natural environment using IndoBERT-based pre-trained models. The flow of our research experiment can be seen in Figure 1. In general, the process is divided into two, namely dataset building as (subsection 3.1) and model building (subsection 3.2).

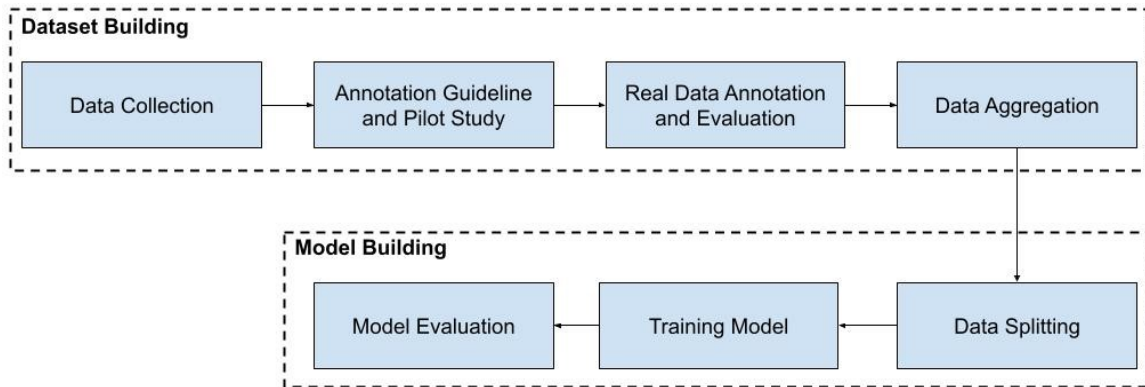


Figure 1. Research experiment flow

#### 3.1. Dataset building

The data used for creating the dataset for this paper comes from Instagram. Data in the form of an Instagram post caption is collected with InstaTouch [31] library within the date range of 9 August 2023 until 10 October 2023. We translate and extend keywords used by [26] and obtained 152 keywords in total [25]. In the data collection process, we use both word/phrase form and hashtag form as done by [26]. From this process, the amount of Instagram data that was successfully collected was 135,007 data (including some duplicate data in which more than one of the keywords occurred).

To get a valid dataset, we perform data filtering before the data annotation process. First, we erased duplicate and non-Indonesian data using Langdetect library [32] to avoid evaluation bias result in the modeling process later. This process gives 37,158 unique Indonesian data. Second, we filtered each data based on whether its content is on ES topics or not to avoid non-ES data contamination in our dataset. Based on preliminary qualitative analysis, there is some data that its content is not on environmental topics but contains the keywords that we used to collect Instagram data. This data instead discusses advertising, promotions, and lifestyle. Therefore, we created a model that can filter Instagram data that has context on environmental topics or not. To the best of our knowledge, there has been no research in Indonesian ES topic classification. We built a dataset for this task with a manually labeled sample of 600 data from previous filtering results with the composition of 300 data for ES topic and 300 data for non-ES topic. Then, we split the sample with a standard split ratio of 80% for training data and 20% for testing data. We used one of IndoBERT-based models namely IndoBERT-base-p1 [24] and achieved a macro F1-score of 87.60%. This result indicates the model is good enough to be used for environmental topic filtering. We do not do many experiments for this process as this is only used for filtering before the manual annotation process. 8,889 data have content on ES topics from this filtering process. Third, we filtered based on automatic sentiment label to avoid extreme data-imbalanced in regards to the manual sentiment labeling process label. We used the Akahana model [33], an IndoBERT-based model that already fine-tuned for Indonesian sentiment classification at the document level. To keep balancedness data across ES topics, we target 500 data for each topic with the composition of 175 with positive sentiment, 150 data with neutral sentiment, and 175 data with negative sentiment from the automatic sentiment labeling process. From 10 ES topics, only five topics achieved this target namely climate change, energy, environment, pollution, and waste, so we obtained 2,500 Instagram posts for the manual annotation process.

For the annotation part, we adapted annotation data guidelines and process from [26] for their English dataset. We annotate expression terms (span-level annotation) which are groups of adjacent or closely connected words in the data that contain a positive expression (labeled as *Exp\_Positive*) or negative expression (labeled as *Exp\_Negative*). The positive expression represents an act of supporting, giving a good assessment, and appreciating an action or policy related to the natural environment. Meanwhile, the nega-

tive expression represents rejecting, giving a bad assessment, and criticizing an action or policy related to the natural environment. We hired three native Indonesian speakers that have at least a bachelor's degree for this task.

The Langing Annotate (an online data annotation platform) [34] is used to annotate the data. To ensure that all three annotators understood the annotation guidelines, we conducted a pilot study by asking them to annotate 150 data samples that have been annotated by us and evaluate the the agreement score between each annotator and with us using pairwise Cohen's  $\kappa$  [29]. If an annotator candidate has an average pairwise Cohen's  $\kappa$  less than 0.4, we will look for new annotator candidate and test them with the same procedure. We follow the procedure of  $\kappa$  interpretation level given by [35] for this examination. Once the average pairwise Cohen's  $\kappa$  for all annotators is more than or equal to 0.4, we ask them to annotate all 2,500 data that we obtained from the filtration process. When the annotators have completed their annotations, we calculated Fleiss'  $\kappa$  [28] to evaluate the agreement level for all annotator results and we got a score of 0.5583, indicating the agreement level has a moderate level [35] and is suitable for the experiment.

The example of annotation results can be seen in Figure 2. In Figure 2(a), there are two terms annotated as negative expressions. The first term "*harus segera dihentikan*" (must be stopped immediately) represents rejection of industry that produces carbon emission. The second term "*berdampak relatif lebih besar terhadap pencemaran lingkungan*" (has a relatively greater impact on environmental pollution) represents criticism of carbon emission produced by industry around the city that leads to environmental pollution. In Figure 2(b) there are two terms annotated as negative expressions and one term annotated as positive expression.

The first term "*bikin cadangan energinya makin sedikit*" (reduce energy reserves) and the second term "*dampak buruknya*" (the bad impact) represent criticism of the impact of continuous use of fossil energy. The third term "*gak akan habis*" (will not run out) represents giving a good assessment of using renewable energy. In Figure 2(c), there are two terms annotated as positive expressions. The first term "*langkah kecil menuju perubahan besar*" (small steps towards big change) and the second term "*mencapai masa depan yang lebih cerah dan berkelanjutan*" (achieve a brighter and more sustainable future) represent an act of supporting the use of renewable energy, smart waste processing, and emission reduction.

In Figure 2(d), there are two terms annotated as positive expressions and one term annotated as negative expression. The first term "*setorkan sampah terpilah ke bank sampah terdekat*" (deposit sorted waste to the nearest waste bank) and the second term "*terintegrasi*" (integrated) represent support and benefits for the use of waste banks in the local community, respectively. The third term "*mencemari lingkungan*" (contaminate the environment) represents the negative impacts that can occur if waste is not managed properly in the local community. In Figure 2(e), there is no term annotated as either negative or positive expression. In Figure 2(f), there are equally terms annotated as positive and negative expressions. The first term "*tenggelam*" (drowned) and "*segala bahayanya*" (all the dangers) are directed to the negative effects about waste felt in the local community. The third term "*Yuk sama-sama*" (let's do it together) and the fourth term "*mengurangi sampah plastik*" (reduce the plastic waste) are directed to the local community to support plastic waste reduction movement.

We use majority voting in code to determine the final label for each data in two steps following [30]. First, it converts span-level labels to document-level labels by performing majority voting on each annotator's results. Majority voting is performed by calculating the number of highest expression terms: labeled as "POSITIVE" if the highest expression term is a positive expression and labeled as "NEGATIVE" if the highest expression term is a negative expression. If the number of terms containing negative expressions and positive expressions is the same or there are no expression terms at all in the data, then the data will be labeled as "NEUTRAL". The sample of conversion result to document-level label can be seen in Table 1. Second, the code determines the final label by performing majority voting on the three document-level labels on each data. This process occurs because we have three document-level labels from three annotator results for each data, but the model we will use requires only one label as input for each data. There is several data is removed because no agreement is found (strong disagreement). Therefore, the total data in the dataset for this experiment is 2,391 data which consists of 1,251 data with the label "POSITIVE", 288 data with the label "NEUTRAL", and 852 data with the label "NEGATIVE".

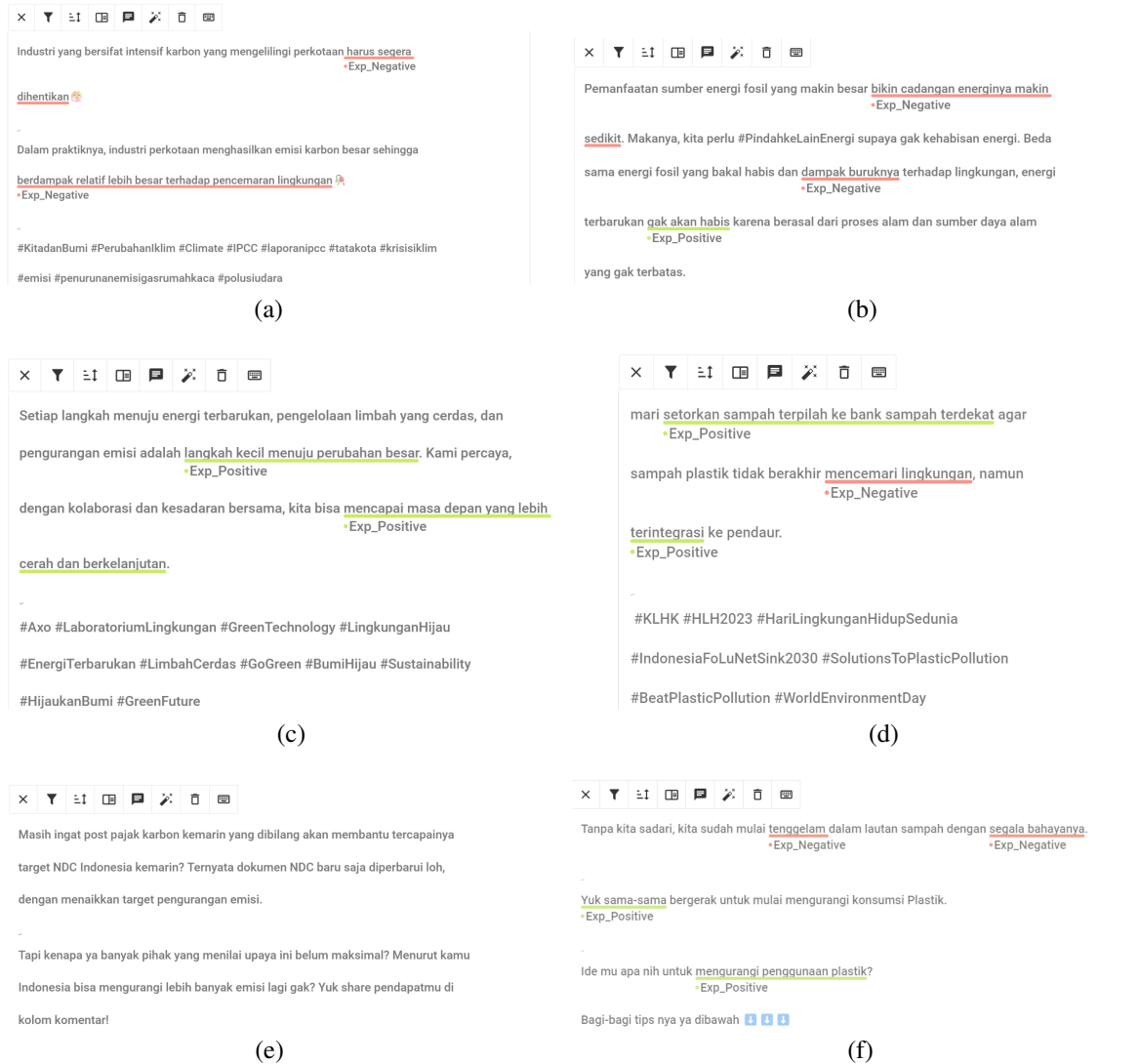


Figure 2. Example of data with span-level labels: (a) negative expression term, (b) majority negative expression term, (c) positive expression term, (d) majority positive expression term, (e) no expression term, and (f) equal expression term

Table 1. Example of document-level label conversion by majority voting on span-level label

Figure	# Positive term	# Negative term	Document-level label
2(a)	0	2	NEGATIVE
2(b)	1	2	NEGATIVE
2(c)	2	0	POSITIVE
2(d)	2	1	POSITIVE
2(e)	0	0	NEUTRAL
2(f)	2	2	NEUTRAL

### 3.2. Model building

After the dataset building process, we split the dataset with a ratio of 80:20 (standard split), 80% of the data will be used as training data and 20% of the data will be used as testing data. We further split the training data with the same ratio, 80% of data will be used as training data and 20% of data will be used as validation data.

We use a variant of IndoBERT models [24] in analyzing sentiment on a dataset that discusses natural environment in Indonesian Instagram, namely IndoBERT-lite-base-p2, IndoBERT-lite-large-p2, IndoBERT-base-p2, and IndoBERT-large-p2. IndoBERT model is a pre-trained language model that was trained on Indonesian corpus Indo4B which is collected from Indonesian Wikipedia, news articles, blogs, and social media, with a size of approximately 4 billion Indonesian words. The model is designed to be fine-tuned on natural language processing tasks such as SA, named entity recognition, question answering, and summarization. IndoBERT is constructed based on two architectures: BERT [20] and ALBERT [36]. IndoBERT-base and IndoBERT-large architectures follow  $BERT_{BASE}$  (12 layers, 768 hidden sizes, 12 self-attention heads, total parameters of 110M) and  $BERT_{LARGE}$  (24 layers, 1,024 hidden sizes, 16 self-attention heads, total parameters of 340 M) architectures respectively. IndoBERT-lite-base and IndoBERT-lite-large architectures follow  $ALBERT_{BASE}$  (12 layers, 768 hidden sizes, 12 self-attention heads, total parameters of 12 M) and  $ALBERT_{LARGE}$  (24 layers, 1,024 hidden sizes, 16 self-attention heads, total parameters of 18M) architectures respectively. Willie *et al.* [24] used two phases for training IndoBERT. In the first phase (p1), IndoBERT was trained on the Indo4B corpus with a maximum sequence length of 128. Then, in the second phase (p2), IndoBERT was further trained on the same corpus with a maximum sequence length of 512. The second phase scheme with a longer sequence allows the model to understand the meaning of words based on the surrounding context more deeply and better. We will use training data to fine-tune all variants of IndoBERT models and validation data to adjust the hyperparameters setup used. Then, We use the fine-tuned IndoBERT model to classify sentiment on testing data to “POSITIVE” (having positive sentiment), “NEGATIVE” (having negative sentiment), or “NEUTRAL” (having neutral sentiment) labels. The simpletransformers [37] library is used to implement all variants of the IndoBERT model.

The evaluation metric used for measuring the performance of the fined-tuned model is macro average F1-score. Macro average F1-score [38] is calculated by averaging the F1-score value in each class with the same weight. The metric is particularly useful in cases of imbalanced datasets, as it provides a balanced evaluation of model performance across all classes. We use Scikit-Learn [39] library to implement the F1-score evaluation.

## 4. RESULTS AND DISCUSSIONS

This section discusses the results of our research including the dataset-building result (subsection 4.1) and model-building results (subsection 4.2).

### 4.1. Data building result and discussion

We built a word cloud to gain insight into what is being discussed based on frequently occurring words in the dataset. Figure 3 shows word cloud per label. Figure 3(a) shows words that often appear in data labeled as “POSITIVE”. Some examples of the words that often appear are “indonesia” (Indonesia), “lingkungan hidup” (environment), “lingkungan” (environment), “kegiatan” (activity), and “sampah” (trash). We concluded that most data labeled as “POSITIVE” discusses environmental activities related to trash in Indonesia. Figure 3(b) shows words that often appear in data labeled as “NEUTRAL”. Some examples of the words that often appear are “sampah” (trash), “indonesia” (Indonesia), “bahan bakar” (fuel), “lingkungan” (environment), and “limbah” (waste). We concluded that most data labeled as “NEUTRAL” discusses waste and fuel in the Indonesian environment. Figure 3(c) shows words that often appear in data labeled as “NEGATIVE”. Some examples of the words that often appear are “indonesia” (Indonesia), “polusi udara” (air pollution), “polusi” (pollution), “lingkungan” (environment), and “sampah” (trash). We concluded that most data labeled as “NEGATIVE” discusses air pollution and trash in the Indonesian environment. Based on the word cloud of the three labels, the frequently occurring words are waste and environment. These word clouds can then be utilized as a support system for the government or stakeholders to better manage the presence of waste in the environment based on public perceptions.

As mentioned in subsection 3.2, for the experiment, we use 20% of data for the test set. From the remaining 80% data, we use 20% of them for the validation set while the rest is for the training set. The dataset distribution of training, validation, and testing data can be seen in Table 2. Table 2 shows that the dataset is quite imbalanced where the positive label becomes the majority label. This imbalanced proportion may give challenges in the modeling process. In this case, if we do not apply the proposed filtering process, we may obtain a more extreme imbalanced dataset.





Table 3. Hyperparameter setup to fine-tune IndoBERT model [20]

Model	Batch size	Epoch	Learning rate
IndoBERT-lite-base-p2	32	15	$4 \times 10^{-5}$
IndoBERT-lite-large-p2	16	15	$1 \times 10^{-5}$
IndoBERT-base-p2	32	15	$4 \times 10^{-5}$
IndoBERT-large-p2	16	15	$4 \times 10^{-5}$

Table 4. Fine-tuned IndoBERT results on testing data

Model	F1-score
IndoBERT-lite-base-p2	45.79%
IndoBERT-lite-large-p2	46.66%
IndoBERT-base-p2	68.49%
IndoBERT-large-p2	72.44%

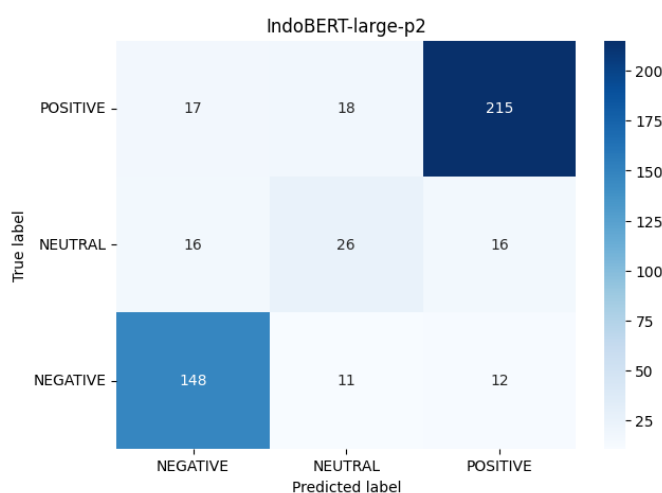


Figure 4. Confusion matrix of model based on IndoBERT-large-p2

For further analysis, we conducted qualitative analysis on a few samples of incorrectly predicted labels. Figure 5 shows sample of incorrectly predicted to have “POSITIVE” label and its annotation results for annotator 1 in Figure 5(a), annotator 2 in Figure 5(b), and annotator 3 in Figure 5(c). Then, Figure 6 shows sample of incorrectly predicted to have “NEGATIVE” label and its annotation results for annotator 1 in Figure 6(a), annotator 2 in Figure 6(b), and annotator 3 in Figure 6(c). Finally, Figure 7 shows sample of incorrectly predicted to have “NEUTRAL” label and its annotation results for annotator 1 in Figure 7(a), annotator 2 in Figure 7(b), and annotator 3 in Figure 7(c). As can be seen in Table 1, the majority voting code we used forces data to be represented by a document-level label even if the data contains multiple expression terms (span-level label). In each of the samples, each annotator result has multiple expression terms (“Exp\_Positive” and “Exp\_Negative”), but no document-level label from the majority voting can represent them well. Therefore, the overall expression terms in data may not be well represented by the final label, so this process could mislead the model in predicting the final label. Instead of directly using a model to predict the final label, it would be interesting to use ensemble learning [43], [44] to determine the final label so it can accommodate multiple expressions terms contained in the data. We can utilize a more granular form of SA, such as sentiment term extraction [45], [46] to identify all expression terms and their labels better as the dataset we have built contains such information. Also, The main focus of Instagram is to convey information through visual component. Therefore, it could be interesting to comprehend sentiment polarity through Instagram visual component and combine it with Instagram text component to gauge public sentiment toward a certain topics (multimodal SA) [47]–[49].

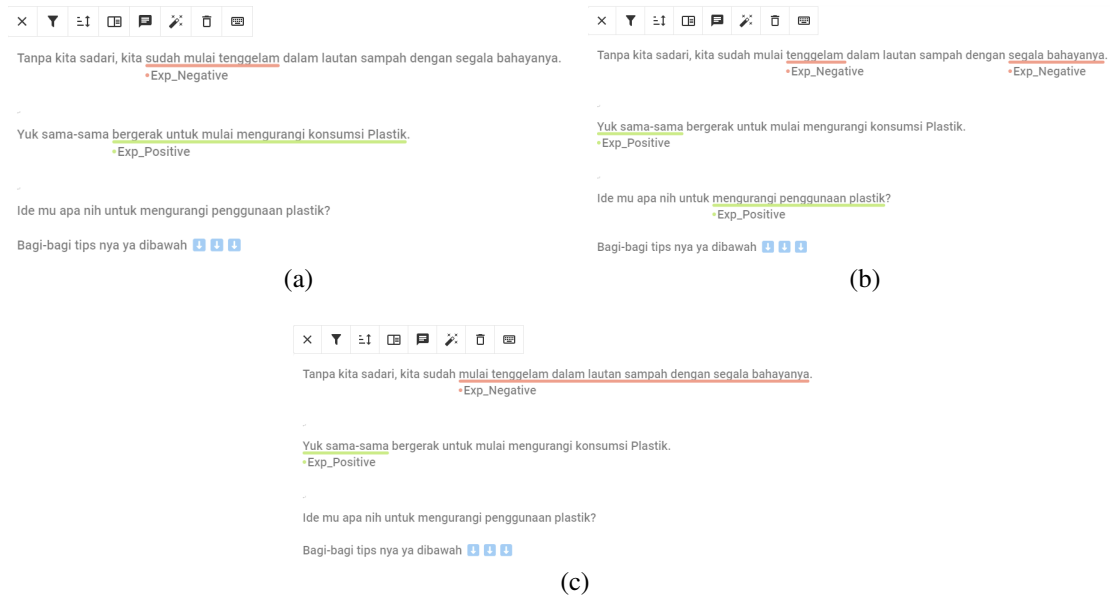


Figure 5. Example of annotator result of data that was incorrectly predicted to have the label “POSITIVE”:  
(a) 1’s, (b) 2’s, and (c) 3’s

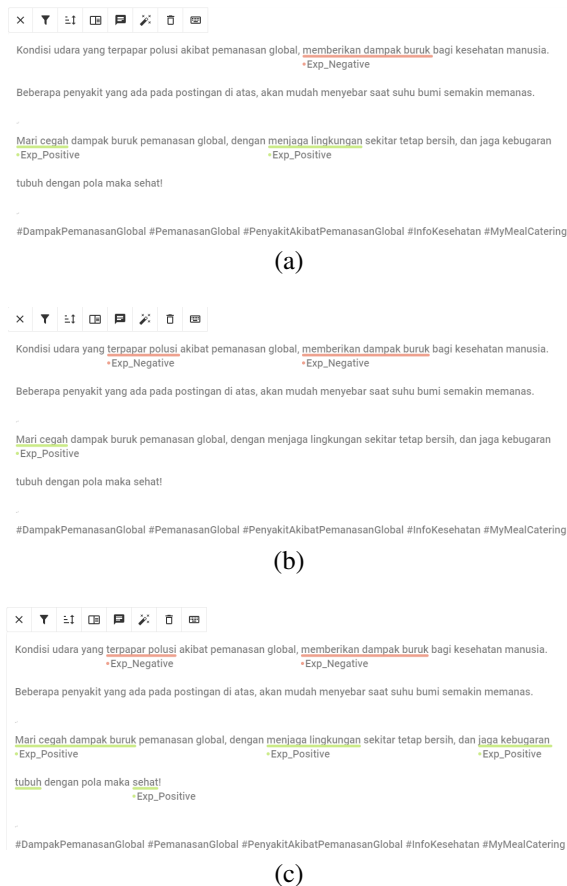


Figure 6. Example of annotator result of data that was incorrectly predicted to have the label “NEGATIVE”:  
(a) 1’s, (b) 2’s, and (c) 3’s

(a)

(b)

(c)

Figure 7. Example of annotator result of data that was incorrectly predicted to have the label “NEUTRAL”:  
 (a) 1’s, (b) 2’s, and (c) 3’s

### 5. CONCLUSIONS AND FUTURE WORKS

In this research, we successfully built a new dataset of Instagram data in the Indonesian language for SA on the natural environment. This dataset can be used by researchers and stakeholders to monitor natural environmental conditions and issues. Hence, the stakeholders can better fit the public’s expectations when

they develop public policies that do not harm the natural environment. The topics of the natural environment we used consist of “Environment”, “Climate Change”, “Waste”, “Energy” and “Pollution”. We used three annotators to manually annotate data and a majority voting method to convert the result into final labels: “POSITIVE”, “NEUTRAL”, and “NEGATIVE”. The size of the dataset is 2,391 consisting of 1,251 data labeled as “POSITIVE”, 288 data labeled as “NEUTRAL”, and 1,251 labeled as “NEGATIVE”. Also, Our dataset not only provides the document-level label, but also gives a more granular label that is a span of expression terms. Before experimenting, We built a word cloud to gain insight into what is being discussed based on the dataset. Based on the word cloud, the frequently occurring words are waste and environment. These results can then be utilized as a support system for the government or stakeholders to better manage the presence of waste in the environment based on public opinions. Then, we experimented with this dataset by comparing the performance of several IndoBERT models. The best model was achieved based on IndoBERT-large-p2 with F1-score of 72.44%. This model has the biggest parameter among other models. We created a confusion matrix to examine the most difficult label to predict. Data with the “NEUTRAL” label was the most difficult label to predict by the model because it naturally has polarity between positive and negative expression. Also, the data used in training was the fewest among other labels. In qualitative error analysis, we concluded that the majority voting process can mislead the model in predicting labels because it can not represent the overall expression terms in data into a final label.

For the future future work in this experiment, we suggest seven improvements. First, the broader topics of the natural environment (i.e. the 10 initial topics we mentioned) could be used to collect more variation about the public’s view toward the natural environment. Second, we could utilize other models with larger parameters since we got the best result from the IndoBERT model with the largest parameters. Third, we can train IndoBERT variants on Instagram data and use it for solving tasks with Instagram datasets better (i.e. IndoBERT-Tweet model). Fourth, increasing the quantity of data labeled as “NEUTRAL” can help improve the model representation since it is the fewest among other labels. Fifth, we can utilize ensemble learning to determine the final label so it can accommodate multiple expressions terms contained in the data. Sixth, we can utilize a more granular form of SA, such as sentiment term extraction to identify all expression terms and their labels better since the dataset is compatible with the task. Seventh, The main focus of Instagram is to convey information through visual component. Therefore, it could be interesting to comprehend sentiment polarity through Instagram visual component and combine it with Instagram text component as features to comprehensively gauge public sentiment toward natural environment topics on Indonesian Instagram (multimodal SA).

## ACKNOWLEDGEMENTS

This work is supported by Hibah Riset Internal Fakultas Ilmu Komputer under a project with grant number NKB-6/UN2.F11.D/HKP.05.00/2024 from Directorate Research and Community Services, Universitas Indonesia.

## REFERENCES




- [1] M. O. Ibrohim, C. Bosco, and V. Basile, “Sentiment analysis for the natural environment: a systematic review,” *ACM Computing Surveys*, vol. 56, no. 4, Nov. 2023, doi: 10.1145/3604605.
- [2] R. G. Rodrigues, R. M. das Dores, C. G. Camilo-Junior, and T. C. Rosa, “SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks,” *International Journal of Medical Informatics*, vol. 85, no. 1, pp. 80–95, Jan. 2016, doi: 10.1016/j.ijmedinf.2015.09.007.
- [3] A. ElMessiry, Z. Zhang, W. O. Cooper, T. F. Catron, J. Karrass, and M. P. Singh, “Leveraging sentiment analysis for classifying patient complaints,” in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Aug. 2017, pp. 44–51, doi: 10.1145/3107411.3107421.
- [4] E. Susilawati, “Public services satisfaction based on sentiment analysis: case study: electrical services in Indonesia,” in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, Oct. 2016, pp. 1–6, doi: 10.1109/ICITSI.2016.7858241.
- [5] F. F. Rachman, R. Nooraeni, and L. Yuliana, “Public opinion of transportation integrated (Jak Lingko), in DKI Jakarta, Indonesia,” *Procedia Computer Science*, vol. 179, pp. 696–703, 2021, doi: 10.1016/j.procs.2021.01.057.
- [6] B. Siswanto, “Sentiment analysis in Indonesian on Jakarta culinary as a recommender system,” in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2021, pp. 46–50, doi: 10.1109/ISRITI54043.2021.9702772.
- [7] M. A. Hadiwijaya, F. P. Pirdaus, D. Andrews, S. Achmad, and R. Sutoyo, “Sentiment analysis on tokopedia product reviews using natural language processing,” in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, Nov. 2023, pp. 380–386, doi: 10.1109/ICIMCIS60089.2023.10348996.

- [8] M. Dahbi, R. Saadane, and S. Mbarki, "Social media sentiment monitoring in smart cities," in *Proceedings of the 4th International Conference on Smart City Applications*, Oct. 2019, pp. 1–6, doi: 10.1145/3368756.3368997.
- [9] D. Zhang *et al.*, "iFoodCloud: a platform for real-time sentiment analysis of public opinion about food safety in China," *arXiv preprint arXiv:2102.11033*, 2021.
- [10] Maskun, H. Assidiq, N. H. Al Mukarramah, and S. N. Bachril, "Threats to the sustainability of biodiversity in Indonesia by the utilization of forest areas for national strategic projects: A normative review," *IOP Conference Series: Earth and Environmental Science*, vol. 886, no. 1, p. 012071, Nov. 2021, doi: 10.1088/1755-1315/886/1/012071.
- [11] C. D. Butler and B. Haryanto, "Climate Change and Health in Indonesia," in *Climate Change and Global Health*, GB: CABI, 2024, pp. 435–444.
- [12] B. Qi, A. Costin, and M. Jia, "A framework with efficient extraction and analysis of Twitter data for evaluating public opinions on transportation services," *Travel Behaviour and Society*, vol. 21, pp. 10–23, 2020, doi: 10.1016/j.tbs.2020.05.005.
- [13] M. R. A. Zein, K. L. Fadillah, N. Febriani, R. Nasrullah, and N. T. Khang, "Social media use for climate change campaign among Indonesian millennials," *PROfesi Humas*, vol. 8, no. 2, p. 168, 2024, doi: 10.24198/prh.v8i2.50167.
- [14] C. Michael and D. N. Utama, "A modified DSM based on social media for treating waste management issue," *ICIC express letters. Part B, Applications: an international journal of research and surveys*, vol. 11, no. 11, pp. 1001–1010, 2020, doi: 10.24507/ici-celb.11.11.1001.
- [15] Z. Indra, A. Setiawan, and Y. Jusman, "Implementation of machine learning for sentiment analysis of social and political orientation in Pekanbaru city," in *Journal of Physics: Conference Series*, 2021, vol. 1803, no. 1, p. 12032, doi: 10.1088/1742-6596/1803/1/012032.
- [16] E. Sugiharti and D. Fauziah, "Comparative study between KNN and maximum entropy classification in sentiment analysis of menstrual cup," in *Journal of Physics: Conference Series*, 2021, vol. 1918, no. 4, p. 42158, doi: 10.1088/1742-6596/1918/4/042158.
- [17] A. Fontanella, Hendrick, H. Ihsan, Z.-H. Wang, and Igra, "Sentiment analysis of Indonesian government policy in the environmental sector using machine learning method," *Review of Accounting, Finance and Governance (RAFGO)*, vol. 1, no. 2, pp. 1–5, 2021.
- [18] B. H. Iswanto and V. Poerwoto, "Sentiment analysis on Bahasa Indonesia tweets using Unigram models and machine learning techniques," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 434, no. 1, p. 12255, doi: 10.1088/1757-899X/434/1/012255.
- [19] A. F. Aji *et al.*, "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," *arXiv preprint arXiv:2203.13357*, 2022.
- [20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [21] Z. A. Guven, "Comparison of BERT models and machine learning methods for sentiment analysis on Turkish Tweets," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, Sep. 2021, pp. 98–101, doi: 10.1109/UBMK52708.2021.9559014.
- [22] E. Demir and M. Bilgin, "Sentiment analysis from Turkish news texts with BERT-based language models and machine learning algorithms," in *2023 8th International Conference on Computer Science and Engineering (UBMK)*, Sep. 2023, pp. 01–04, doi: 10.1109/UBMK59864.2023.10286719.
- [23] A. J. Nair, V. G., and A. Vinayak, "Comparative study of Twitter sentiment on COVID - 19 Tweets," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 1773–1778, doi: 10.1109/IC-CMC51019.2021.9418320.
- [24] B. Willie *et al.*, "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.
- [25] C. Octovianto, "Indonesian sentiment analysis in natural environment topic," *GitHub*. <https://github.com/christoferoctovianto/id-sa-in-ne> (accessed Feb. 05, 2025).
- [26] C. Bosco, M. O. Ibrohim, V. Basile, and I. Budi, "How green is sentiment analysis? Environmental topics in corpora at the University of Turin," in *CEUR Workshop Proceedings*, 2023, vol. 3596.
- [27] "Prolific." <https://www.prolific.com/> (accessed Feb. 05, 2025).
- [28] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971, doi: 10.1037/h0031619.
- [29] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [30] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti, "Overview of the Evalita 2016 sentiment polarity classification task," in *EVALITA. Evaluation of NLP and Speech Tools for Italian*, vol. 1749, Torino: Accademia University Press, 2016, pp. 146–155.
- [31] A. Nord and Esteban, "Instagram-scraper," *GitHub*. <https://github.com/drawrowfly/instagram-scraper> (accessed Feb. 05, 2025).
- [32] M. Danilk, "langdetect 1.0.9," *pypi.org*, 2021. <https://pypi.org/project/langdetect/> (accessed Feb. 05, 2025).
- [33] F. Alfarid, "Indonesia-sentiment-roberta," *Hugging Face*, 2021. <https://huggingface.co/akahana/indonesia-sentiment-roberta> (accessed Feb. 05, 2025).
- [34] "Langing anotate," *langing.ai*. 2023. [Online]. Available: <https://beta.annotate.langing.ai/> (accessed Feb. 05, 2025).
- [35] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977, doi: 10.2307/2529310.
- [36] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soiccut, "Albert: a lite BERT for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2020.
- [37] T. Rajapakse, "Simple transformers," *GitHub*, 2021. <https://github.com/ThilinaRajapakse/simpletransformers> (accessed Feb. 05, 2025).




- [38] A. J. Nair, V. G., and A. Vinayak, "Comparative study of Twitter sentiment on COVID - 19 Tweets," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 1773–1778, doi: 10.1109/ICCMC51019.2021.9418320.
- [39] "F1\_score," *Scikit Learn*. [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html) (accessed Feb. 05, 2025).
- [40] 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*, Nov. 2021, pp. 10660–10668, doi: 10.18653/v1/2021.emnlp-main.833.
- [41] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 26–33, doi: 10.3115/1073012.1073017.
- [42] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009, doi: 10.1109/MIS.2009.36.
- [43] O. Gharroudi, H. Elghazel, and A. Aussem, "Ensemble multi-label classification: a comparative study on threshold selection and voting methods," in *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2015, vol. 2016-Janua, pp. 377–384, doi: 10.1109/ICTAI.2015.64.
- [44] A. Mahdavi-Shahri, M. Houshmand, M. Yaghoobi, and M. Jalali, "Applying an ensemble learning method for improving multi-label classification performance," in *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, Dec. 2016, pp. 1–6, doi: 10.1109/ICSPIS.2016.7869900.
- [45] J. Fernando, M. L. Khodra, and A. A. Septiandri, "Aspect and opinion terms extraction using double embeddings and attention mechanism for Indonesian hotel reviews," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Sep. 2019, pp. 1–6, doi: 10.1109/ICAICTA.2019.8904124.
- [46] Y. A. Winatmoko, A. A. Septiandri, and A. P. Sutiono, "Aspect and opinion term extraction for hotel reviews using transfer learning and auxiliary labels," *arXiv preprint arXiv:1909.11879*, 2019, [Online]. Available: <http://arxiv.org/abs/1909.11879>.
- [47] C. Li and Z. Hu, "Multimodal sentiment analysis of social media based on top-layer fusion," in *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, Dec. 2022, pp. 1–6, doi: 10.1109/ICCC56324.2022.10065604.
- [48] S. Mekala, M. Bhuvana, D. Brat Gupta, V. Bhatt, P. Kunekar, and G. Manoharan, "Natural language processing and deep learning techniques to improve sentiment analysis in social media texts," in *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023*, 2023, vol. 6, pp. 1751–1755, doi: 10.1109/IC3I59117.2023.10397735.
- [49] S. Kapoor, S. Gulati, S. Verma, A. Pandey, and D. K. Vishwakarma, "Multimodal architecture for sentiment recognition via employing multiple modalities," in *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, May 2024, pp. 435–439, doi: 10.1109/InCACCT61598.2024.10551131.

## BIOGRAPHIES OF AUTHORS






**Christofer Octovianto**    holds a bachelor's degree in Engineering Physics (B.Eng) from Institut Teknologi Bandung. While studying in college, he was active doing activities in Automation and Control Lab. Currently, he is master student of computer science at Universitas Indonesia under the supervision of Prof. Dr. Indra Budi and Muhammad Okky Ibrohim, M.Kom. His research interests are related to AI and natural language processing. He can be contacted at email: christofer.octovianto@cs.ui.ac.id.



**Muhammad Okky Ibrohim**    is a Ph.D. student at the Dipartimento di Informatica, Università Degli Studi di Torino, Italy. Before taking a Ph.D., he is a lecturer and researcher at the Faculty of Computer Science, Universitas Indonesia, Indonesia. His research interests are focused on natural language processing for social good, including sentiment analysis about environmental sustainability, hate speech detection, and abusive language detection in general. He can be contacted at email: muhammadokky.ibrohim@unito.it.



**Indra Budi**    is a full professor in the Faculty of Computer Science, Universitas Indonesia, Indonesia. He took the bachelor's, master's, and Ph.D. in computer science at the same university he works now. He has been teaching at the university since July 2021. His research interests include information retrieval, text and data mining, and other natural language processing topics in general. He can be contacted at email: indra@cs.ui.ac.id.