Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemendiktisaintek, No. 10/C/C3/DT.05.00/2025

Automatic Analysis of Natural Disaster Messages on Social Media using IndoBERT and Multilingual BERT

Yasmin Dwi Safitri[1], Mohammad Reza Faisal [2], Dwi Kartini[3], Triando Hamonangan Saragih[4], Friska Abadi[5], Adam Mukharil Bachtiar[6]

[1,2,3,4,5] Department of Computer Science, Faculty of Mathematics and Natural Science,
[6]School of Knowledge Science
[1,2,3,4,5] Lambung Mangkurat University, Banjarbaru, Indonesia
[6]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: 2111016220005@mhs.ulm.ac.id[1], reza.faisal@ulm.ac.id[2], dwikartini@ulm.ac.id[3], triando.saragih@ulm.ac.id[4], friska.abadi@ulm.ac.id[5], adam.bachtiar@jaist.ac.jp[6]

ARTICLE INFO

History of the article:
Received June 17, 2025
Revised August 24, 2025
Accepted August 30, 2025

Keywords:
Deep Learning
Social Media
Natural Disaster
IndoBERT;
Multilingual BERT

Correspondence:
E-mail:
reza.faisal@ulm.ac.id

ABSTRACT

Information about natural disasters disseminated through social media can serve as an important data source for mitigation processes and early warning systems. Social media platforms, such as X (formerly known as Twitter), have become primary channels for conveying real-time information, especially during disaster emergencies. With the large amount of unstructured disaster-related text that must be processed, the main challenge is accurately filtering and classifying messages into three categories: eyewitness, non-eyewitness, and don't know. This research aims to compare the performance of four BERT-based natural language processing models, namely IndoBERT, IndoBERT with Masked Language Modeling (MLM), Multilingual BERT, and Multilingual BERT with MLM, in classifying Indonesian-language disaster messages. The dataset used in this study was obtained from previous research and publicly available data on GitHub, consisting of annotated messages related to floods, earthquakes, and forest fires. The method applied is a deep learning approach using the hold-out technique with an 80:20 ratio for training and testing data, and the same ratio applied to split the training data into training and validation subsets, with stratification to maintain balanced class proportions. In addition, variations in batch size were explored to evaluate their effect on model performance stability. The results show that the IndoBERT model achieved the highest performance on the flood and earthquake datasets, with accuracies of 80.67% and 81.50%, respectively. Meanwhile, IndoBERT with MLM pre-training recorded the highest accuracy on the forest fire dataset, 88.33%. Overall, IndoBERT demonstrated the most consistent and superior performance across datasets compared to the other models. These findings indicate that IndoBERT has strong capabilities in understanding Indonesian disaster-related text, and the results can be used as a foundation for developing automatic classification systems to support real-time disaster monitoring and early warning applications.

INTRODUCTION

Indonesia is among the most disaster-prone countries globally due to its geotectonic positioning at the confluence of three major tectonic plates and its inclusion within the Pacific Ring of Fire (Fuady et al., 2021). This tectonic intersection induces persistent seismic and volcanic activity, rendering the nation highly susceptible to natural hazards such as floods, earthquakes, and forest fires. The high recurrence and intensity of these hazards impose substantial challenges on disaster risk reduction and emergency response

systems. Their impacts are far-reaching, encompassing socioeconomic disruption, environmental degradation, and infrastructure loss (Shidik et al., 2024). Consequently, the establishment of robust early-warning and rapid detection mechanisms becomes imperative to attenuate cascading impacts and accelerate mitigation workflows.

Information regarding natural disasters is spreading rapidly through social media, which plays an essential role in various phases of disaster management. One of the most widely used platforms in this context is Twitter, now known as X, which allows users to share information related to disaster events actively (Amriza et al., 2022). Social media X has a wide user base from various groups and supports the rapid dissemination of information in text form (Komara & Hadiapurwa, 2022). These characteristics make X not only important in the context of information dissemination but also attract the attention of researchers in the field of sentiment analysis, which is part of text mining, due to the high volume and variety of textual data available (Wang et al., 2022). This can prove that text data from X can be used in text mining.

In conventional text mining workflows, feature representation and classification are typically treated as separate stages, with techniques such as TF–IDF commonly employed for feature extraction prior to model training. The advent of BERT introduced a paradigm shift by integrating contextual representation learning and classification within a unified transformer architecture, thereby minimizing the reliance on manual feature engineering (Garrido-Merchan et al., 2023; Gasmi, 2022). BERT, as a bidirectional transformer language model, generates deep contextual embeddings that can be effectively transferred to diverse natural language processing (NLP) tasks (Koroteev MV, 2021). BERT developed its variant trained on 104 languages called Multilingual BERT (Khan et al., 2022). With that advantage, Multilingual BERT can learn good multilingual representations with strong cross-lingual zero-shot performance in various tasks (Wu, 2022). However, in previous journals (Tanvir et al., 2021), Multilingual BERT performed less than the monolingual model (EstBERT) in 5 out of 7 Estonian NLP tasks. This shows that Multilingual BERT has limitations in deepening the understanding of specific languages because the model's capacity is divided into more than 100 languages. In addition, the BERT model was also developed for the monolingual Indonesian BERT model named IndoBERT (Koto et al., 2020). Unlike Multilingual BERT, the IndoBERT model is trained only with Indonesian corpora (Pradnyana et al., 2023). Comparative studies further confirm IndoBERT's advantage: Mahardika et al. (2023) reported that IndoBERT achieved 88% accuracy in hate speech classification on Indonesian tweets, outperforming Multilingual BERT, which only reached 77%. However, IndoBERT still has a drawback in that it lacks the ability to capture linguistic nuances from Indonesia's diverse regional languages.

A substantial body of prior work has explored BERT and its derivatives for Indonesian text classification. Noor Fakhruzzaman et al. (2021) achieved 91.4% accuracy in detecting clickbait headlines in Indonesian using Multilingual BERT. Khan et al. (2022) attained an F1-score of 81.49% for Urdu sentiment analysis using Multilingual BERT. Aygun et al. (2022) conducted aspect-based sentiment analysis on multilingual COVID-19 vaccination tweets, reporting 86% accuracy using Multilingual BERT for English texts. These studies affirm Multilingual BERT's robustness in multilingual and cross-domain contexts.

Other research underscores IndoBERT's superior alignment with Indonesian texts. Nabiilah et al. (2022) attained an F1-score of 88.97% on toxic comment classification using IndoBERT. Ingkafi et al. (2023) reported 86% accuracy in public sentiment analysis on Indonesian vaccination programs. Uliniansyah et al. (2024) achieved 82.63% accuracy in analyzing biodiversity policy discourse. Pramana et

al. (2024) employed IndoBERT for emotion analysis in product reviews with an F1-score of 75%. In addition, a study by Nabiilah & Suhartono (2023) compared IndoBERT, mBERT, and Indonesian RoBERTa for personality classification, obtaining 73.72% accuracy for both IndoBERT and mBERT. Collectively, these findings highlight IndoBERT's consistent performance advantages in Indonesian-specific NLP tasks.

Parallel research on the same disaster-related datasets used in this study employed convolutional neural network (CNN) architectures. Faisal et al. (2022) reported highest accuracies of 78.33% (flood), 78.33% (earthquake), and 81.97% (forest fire) using 2D CNNs. Delimayanti et al. (2020) achieved 77.87% accuracy on the flood dataset using support vector machines (SVM). Nooralifa et al. (2021) attained 64.43% accuracy on the earthquake dataset with SVM, and Rinaldi et al. (2021) replicated the 81.97% CNN result on the forest fire dataset. However, transformer-based architectures such as IndoBERT and Multilingual BERT have yet to be systematically applied to these datasets, presenting a gap for further exploration.

Research conducted by Joshi et al. (2020) highlighted that Masked Language Modeling (MLM) is an important component in language models such as BERT because it can improve performance compared to models without MLM pretraining. However, few studies have explicitly compared the performance of models with and without MLM in the context of the Indonesian language. Therefore, this research analyzes the performance differences between IndoBERT and Multilingual BERT with and without MLM to classify Indonesian-language disaster messages from social media.

Synthesizing insights from the literature, IndoBERT and Multilingual BERT offer complementary strengths: IndoBERT provides domain-specific linguistic depth, whereas Multilingual BERT offers multilingual generalization. MLM pre-training has the potential to augment both, yet its practical effect in Indonesian disaster communication remains empirically underexplored. The novelty of this work lies in evaluating the interaction between MLM pre-training and model architecture (monolingual vs. multilingual) on disaster-related text classification.

Accordingly, this research aims to compare the performance of four transformer-based models (1) IndoBERT, (2) IndoBERT with MLM pre-training, (3) Multilingual BERT, and (4) Multilingual BERT with MLM pre-training in classifying Indonesian disaster messages. A deep learning approach is employed with systematic variation of batch size to examine its influence on model performance. The contributions of this study are threefold: (1) a comparative evaluation of IndoBERT and Multilingual BERT with and without MLM pre-training for Indonesian disaster-text classification; (2) providing empirical insights on the impact of batch size variation toward classification performance in terms of accuracy; and (3) offering recommendations for developing more effective NLP-based systems to support disaster monitoring and early warning applications in Indonesia.

RESEARCH METHODS

To classify text in the dataset related to disasters, such as floods, earthquakes, and forest fires, this study uses pre-trained language models, IndoBERT and Multilingual BERT. This study also examines how these models are trained using the Masked Language Model (MLM) method before classification. Figure 1 shows the research procedure using the pre-trained model. The process begins with preparing the dataset, followed by text normalization to clean and standardize the input. After that, special tokens such as [CLS], [SEP], and [PAD] are added as required by the BERT architecture. The dataset is then split into training and testing subsets, with the same ratio for the validation split on the training data. Fine-tuning is carried

out on the IndoBERT and Multilingual BERT models using the labeled data, and the final stage is evaluation based on accuracy, sensitivity, and specificity.

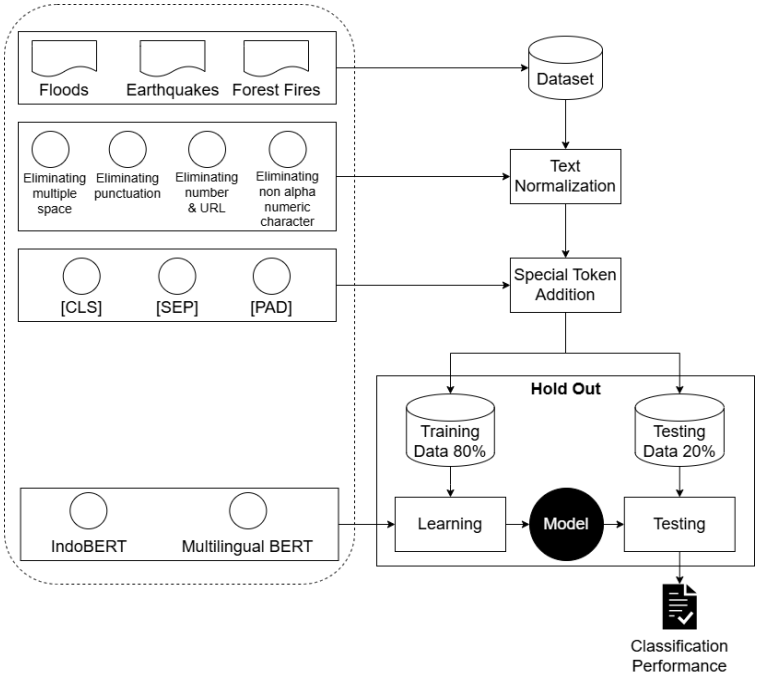

Figure 1. Research procedures with pre-trained models

The study also examines a different approach. Figure 2 illustrates the workflow with an additional Masked Language Modeling (MLM) stage. After dataset preparation, normalization, and special token addition, the models undergo MLM pretraining using the available text data. The pretrained models are then fine-tuned on the labeled dataset in the same way as Figure 1. Finally, evaluation is conducted to measure model performance. This approach is expected to enhance the model's ability to capture contextual patterns in Indonesian disaster messages.

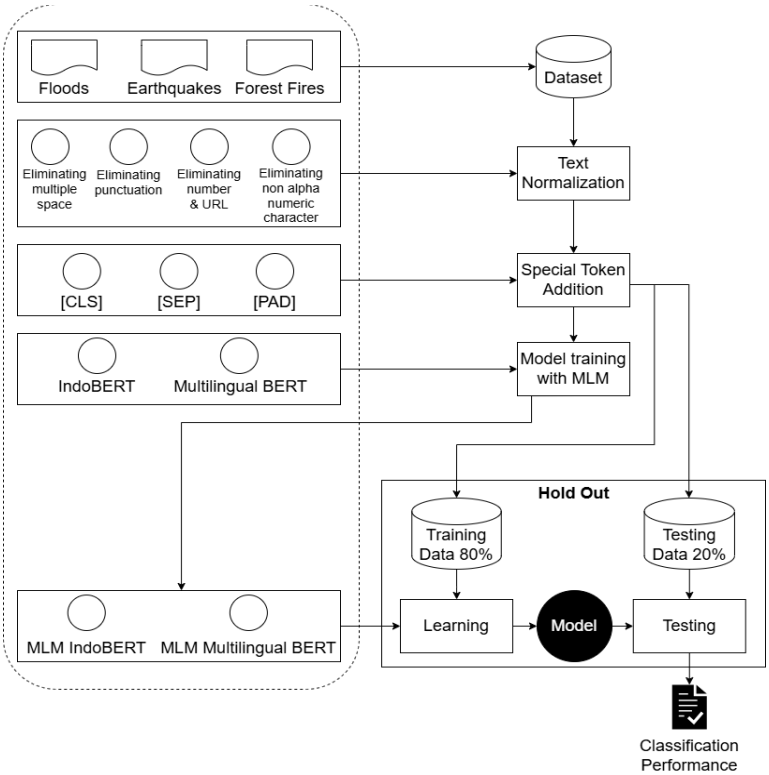

Figure 2. Research procedure with a training model using MLM

1. Dataset

The dataset used in this study is a collection of natural disaster message reports from social media that have been used in previous research (Faisal et al., 2022). All data in the dataset was obtained from social media X through a crawling process that used tweet keyword filters such as "banjir" (floods), "gempa bumi" (earthquakes), "kebakaran hutan" (forest fires), "kebakaran" (fires), "asap" (smoke), and others. All tweets in this dataset are in Indonesian and have been manually labeled to ensure that the tweets match their respective categories. An early warning system for natural disasters can utilize real-time data, which makes the selection of this dataset relevant. The dataset details in this study are presented in Table 1, and examples of natural disaster messages are presented in Table 2.

Table 1. Dataset

Dataset	*Class Label*			*Total*
	Eyewitness	*Noneyewitness*	*Don'tknow*	
Floods	1000	1000	1000	3000
Earthquakes	1000	1000	1000	3000
Forest Fires	1000	1000	1000	3000

Table 2. Example of a natural disaster message

Dataset	*Messages*	*Class Label*
Floods	depan rumah dh banjir huhu	eyewitness
	fenomena banjir sering terjadi di perkotaan	noneyewitness
	hati ku banjir bang	don'tknow
Earthquakes	pantesan goyang goyang gempa ternyata	eyewitness
	waspada gempa dalam waktu dekat ini	noneyewitness
	santai studio kita anti gempa aman	don'tknow
Forest Fires	kirain ngawur lihat jendela beneran kabut ternyata asap kebakaran hutan kah atau gimana nih	eyewitness
	personil polsek gambut sampaikan bahaya kebakaran hutan dan asap bagi kesehatan	noneyewitness
	eh buset kyk org kebakaran jenggot nih orang biasa aja cuy	don’tknow

2. Text normalization

Text normalization is a stage for converting text into a standard form to prepare it, or the following stages training the model to be applied (S. Kumar, 2024). The text normalization carried out in this study is data cleaning, which means cleaning the text data by removing double spaces, punctuation marks, numbers, URLs, as well as non-alphanumeric characters, and case folding, which is standardizing the letter case into lowercase or uppercase (Rahman Isnain et al., 2021). In this study, all text data was standardized into lowercase.

3. Special token addition

BERT employs the WordPiece tokenization strategy, which decomposes words into smaller sub-word units to handle large vocabularies and out-of-vocabulary terms more effectively (Guo et al., 2022). After tokenization, the input text must be represented as a sequence of tokens that conforms to the requirements of the BERT architecture. One important step in this process is the addition of special tokens. In this study, each input sequence was standardized to a maximum length of 128 tokens. If the tokenized sentence is shorter than this limit, additional tokens are appended to reach 128, while longer sequences are truncated to fit the defined length. The [CLS] token is placed at the beginning of each sequence and serves as an aggregate representation of the entire input, where its final hidden state is utilized for classification tasks. At the end of the sequence, the [SEP] token is added to indicate the

boundary or the end of a sentence, which is particularly useful in distinguishing separate textual segments. Meanwhile, the [PAD] token is employed to fill sequences that fall below the maximum length, ensuring that all inputs within a batch have a uniform size of 128 tokens. Through the consistent application of [CLS], [SEP], and [PAD], the representation of disaster-related messages adheres to the BERT input format and supports efficient computation during training, while the choice of maximum length 128 strikes a balance between computational efficiency and capturing adequate contextual information from social media texts (Faisal et al., 2025).

4. Training model with MLM

Masked Language Model (MLM) is a model that predicts masked words, which is not only effective for pretraining but also serves as an efficient data augmentation technique and helps improve model performance in downstream NLP tasks. During training, most of the input data is masked or hidden, with the goal that masked language modeling allows BERT to predict the masked words based on the context provided by the unmasked words. Unlike traditional language models, the MLM approach in BERT enables the model to learn deep bidirectional representations of sentences. This enhances the model's understanding of context and semantic relationships between words, significantly improving various language-processing tasks (Ma, 2023).

In implementing the BERT model, 15% of the total data tokens are taken as candidate tokens to be replaced with the [MASK] token. Then, of all the candidate tokens, 80% of the tokens will be replaced with [MASK], 10% will be replaced with random tokens based on unigram distribution, and the remaining 10% will not be replaced or left as is. The task of this approach is to predict the original token from the replaced token (Joshi et al., 2020). BERT chooses a masking rate of 15% because the model cannot learn good representations when too much text is masked, and training becomes inefficient when too little is masked (Wettig et al., 2022).

MLM works with a principle similar to sentence completion: masking words in the input using the [MASK] token, then predicting those words after the language model is trained, as shown in Figure 3. MLM is a neural network model that functions to predict words that are deliberately hidden in a sentence. By using the MLM approach, the translation system can better understand the context of the source sentence and improve translation quality in terms of accuracy and fluency. This model works bidirectionally, utilizing information before and after the [MASK] token, and does not follow a specific direction (non-directional) (Yang & Yang, 2025).

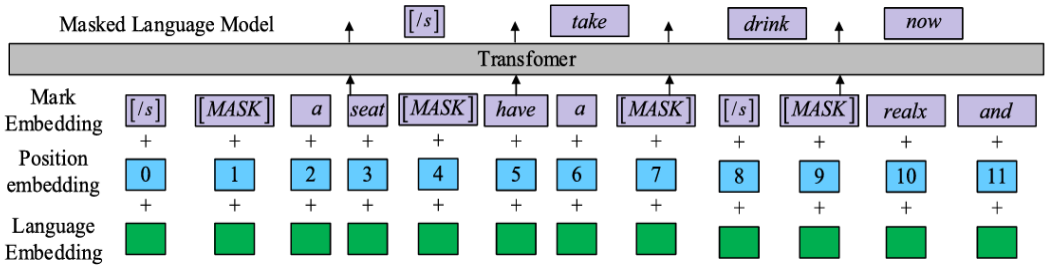

Figure 3. MLM Structure (Yang & Yang, 2025)

5. Classification

This study uses the hold-out method by dividing 80% of the training data and 20% of the testing data, with the same ratio for the validation split on the training data. Stratification is also applied to maintain a balanced class distribution in each data subset (Sadaiyandi et al., 2023), and this can

improve the evaluation's reliability and minimize the model's potential bias toward dominant classes. The training data will be used to build classification models using (1) pre-trained IndoBERT, (2) pre-trained Multilingual BERT, (3) IndoBERT with MLM, and (4) Multilingual BERT with MLM.

Considering that IndoBERT and Multilingual BERT are developments of the BERT architecture, both have the same basic structure as the original BERT. BERT has become one of the most popular NLP models because of its ability to overcome various challenges, such as understanding out-of-vocabulary words through subword tokenization and understanding context bidirectionally. These advantages make BERT very effective in handling text classification tasks, including disaster-related messages, where the meaning of a word highly depends on the overall context of the sentence.

Architecturally, BERT is built with a stack of encoders from the transformer model, consisting of 12 layers (BERTbase) and 24 layers (BERTlarge). Figure 4 shows the number of encoders in the BERT model's encoder stack. Each encoder has two main components: a multi-head self-attention mechanism to capture relationships between words in a sentence and a feed-forward network to detect complex patterns not captured by attention. For pretraining, BERT uses two main approaches: Masked Language Modeling (MLM), which predicts hidden tokens in a bidirectional context, and Next Sentence Prediction (NSP), to understand the relationship between sentence segments. These two techniques allow BERT to understand sentence structure and meaning deeply and to handle new words adaptively.

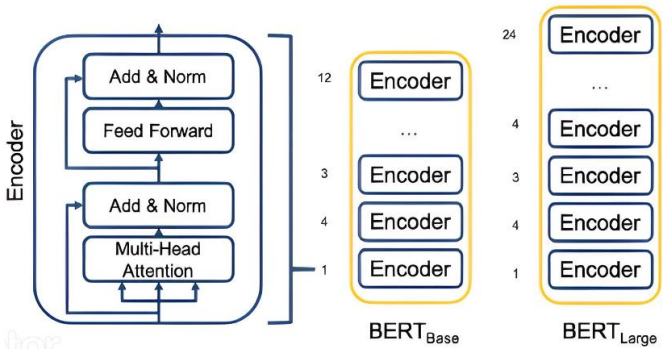

Figure 4. BERT model architecture (K. A. Kumar & Renuka, 2024)

In this study, two BERT variants are used to support the classification of Indonesian-language text: IndoBERT and Multilingual BERT. IndoBERT is a monolingual model specifically trained using Indonesian language corpora, while Multilingual BERT is trained in more than 100 languages, including Indonesian. Although both have a base architecture similar to the original BERT, Multilingual BERT has several adjustments to handle linguistic diversity. At the same time, IndoBERT is optimized to understand the linguistic characteristics of the Indonesian language more deeply (Mahardika et al., 2023).

6. Evaluation

In this study, classification performance evaluation was carried out to assess how well the model can classify disaster-related messages such as floods, earthquakes, and forest fires into three classes, namely eyewitness, non-eyewitness, and don't know. The confusion matrix was used in the evaluation process to calculate the metrics of accuracy, sensitivity, and specificity to measure the quality of predictions produced by the model.

Table 3. Confusion matrix multiclass

Classes		*True Class*		
		A	*B*	*C*
Predicted Class	*A*	TP_A	E_{BA}	E_{CA}
	B	E_{AB}	TP_B	E_{CB}
	C	E_{AC}	E_{BC}	TP_C

Table 3 shows the confusion matrix for multiclass classification with three classes (A, B, and C). There are four outputs from the confusion matrix, namely True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) (Tharwat, 2021). Equations (1), (2), and (3) show the formulas for accuracy, sensitivity, and specificity in multiclass classification.

$$Accuracy = \frac{TP_A + TP_B + TP_C + \ldots + TP_n}{Overall\ sample\ total} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{FP+TN} \quad (3)$$

7. Experimental Setup

This section explains the parameters used in the research process. Four models were used in text classification, IndoBERT and Multilingual BERT, each with and without retraining using Masked Language Modeling (MLM). In Table 4, models labeled "Yes" in the MLM column indicate that the model has undergone additional pre-training using the MLM technique. All models were trained with the same parameters, except for the batch size, which was varied to observe its effect on model performance. In this research, other parameters such as optimizer, learning rate, number of epochs, and maximum input length were kept constant throughout the experiment. Detailed information on the configuration of each model is presented in Table 4.

Table 4. Parameter model classification

No.	*MLM*	*Classification Model*	*Parameter*
1	No	IndoBERT	Optimizer = Adam Learning Rate = 2e-5 Batch Size = [16, 32, 64, 128, 256] Epoch = 10 Max Length = 128
2	Yes		
3	No	Multilingual BERT	
4	Yes		

RESULTS AND DISCUSSION

1. Results

The text normalization stage on the dataset has already been carried out in previous research, so this section proceeds directly to adding special tokens before classification. Tokens such as [CLS] are added at the beginning of each sentence to represent the overall meaning of the input in the classification task, while [SEP] is used as a marker for the end of the sentence. In addition, [PAD] is inserted at the end of text whose length is less than the maximum limit to ensure uniformity of input length. Each of these tokens is then converted into a numerical representation based on its index in the model's vocabulary. This process is illustrated in Table 5, which shows an example of adding special tokens to the input text.

Table 5. Results of adding special tokens

IndoBERT	*Multilingual BERT*
['[CLS]', 'banjir', 'terp', '##arah', 'dari', 'sema', '##lem', 'sampe', 'skrg', 'uj', '##an', 'belum', 'ber', '##enti', 'ga', 'kebayang', 'yg', 'rumahnya', 'biasa', 'banjir', '[SEP]', '[PAD]',...,'[PAD]']	['[CLS]', 'ban', '##jir', 'ter', '##para', '##h', 'dari', 'sem', '##ale', '##m', 'sam', '##pe', 'sk', '##rg', 'uj', '##an', 'belum', 'bere', '##nti', 'ga', 'ke', '##bay', '##ang', 'yg', 'rumah', '##nya', 'biasa', 'ban', '##jir', '[SEP]', '[PAD]',..., '[PAD]']

In the BERT model process, the dataset in the form of tokens will first undergo encoding and an attention mask before entering the classification stage. At the classification stage, the performance of the classification models is tested using three types of disasters, namely floods, earthquakes, and forest fires, as summarized in Table 6. The evaluation was conducted on four models: IndoBERT, IndoBERT MLM, Multilingual BERT, and Multilingual BERT MLM. Each model was evaluated based on accuracy, sensitivity, and specificity and tested using various batch sizes to observe their effect on classification performance.

Table 6. Classification model results

Classification Models	*Dataset*	*Batch Size*	*Classification Performance*		
			Accuracy (%)	*Sensitivity (%)*	*Specificity (%)*
IndoBERT	Floods	16	**80.67**	**80.67**	**90.33**
		32	79.33	79.33	89.67
		64	73.33	73.33	86.67
		128	80.50	80.50	90.25
		256	**80.67**	**80.67**	**90.33**
	Earthquakes	16	79.17	79.17	89.58
		32	78.50	78.50	89.25
		64	81.33	81.33	90.67
		128	80.67	80.67	90.33
		256	**81.50**	**81.50**	**90.75**
	Forest Fires	16	87.50	87.50	93.75
		32	84.67	84.67	92.33
		64	**88.17**	**88.17**	**94.08**
		128	87.67	87.67	93.83
		256	86.33	86.33	93.17
IndoBERT MLM	Floods	16	77.67	77.67	88.83
		32	78	78	89
		64	78.67	78.67	89.33
		128	**78.83**	**78.83**	**89.42**
		256	77.83	77.83	88.92
	Earthquakes	16	78.83	78.83	89.42
		32	**80.33**	**80.33**	**90.17**
		64	79.50	79.50	89.75
		128	77.33	77.33	88.67
		256	79.50	79.50	89.75
	Forest Fires	16	**88.33**	**88.33**	**94.17**
		32	87.17	87.17	93.58
		64	85.17	85.17	92.58
		128	87.83	87.83	93.92
		256	87.33	87.33	93.67
Multilingual BERT	Floods	16	**77.67**	**77.67**	**88.83**
		32	76.50	76.50	88.25
		64	74.33	74.33	87.17
		128	76.17	76.17	88.08
		256	73	73	86.50
	Earthquakes	16	**78.67**	**78.67**	**89.33**
		32	77.83	77.83	88.92
		64	77.17	77.17	88.58
		128	78.17	78.17	89.08
		256	74.67	74.67	87.33

Classification Models	*Dataset*	*Batch Size*	*Classification Performance*		
			Accuracy (%)	*Sensitivity (%)*	*Specificity (%)*
Multilingual BERT MLM	Forest Fires	16	86	86	93
		32	86.33	86.33	93.17
		64	85	85	92.50
		128	86.33	86.33	93.17
		256	**86.50**	**86.50**	**93.25**
	Floods	16	75.83	75.83	87.92
		32	75.17	75.17	87.58
		64	73.17	73.17	86.58
		128	74.67	74.67	87.33
		256	**76.67**	**76.67**	**88.33**
	Earthquakes	16	75.33	75.33	87.67
		32	**78**	**78**	**89**
		64	76	76	88
		128	76.83	76.83	88.42
		256	77.67	77.67	88.83
	Forest Fires	16	85.67	85.67	92.83
		32	86.17	86.17	93.08
		64	84.50	84.50	92.25
		128	87.50	87.50	93.75
		256	**88**	**88**	**94**

The best classification model performance on the flood dataset was obtained using the IndoBERT model with batch sizes 16 and 256, which produced an accuracy of 80.67%, sensitivity of 80.67%, and specificity of 90.33%. Like the flood dataset, the earthquake dataset also achieved the best performance using the IndoBERT model with a batch size of 256, which produced an accuracy of 81.50%, sensitivity of 81.50%, and specificity of 90.75%. In contrast, the forest fire dataset achieved its best performance using the IndoBERT MLM model with batch size 16, which produced an accuracy of 88.33%, sensitivity of 88.33%, and specificity of 94.17%.

2. Discussion

After obtaining classification results from various model configurations, datasets, and batch sizes, this stage further discusses the performance of each model based on the evaluation metrics used.

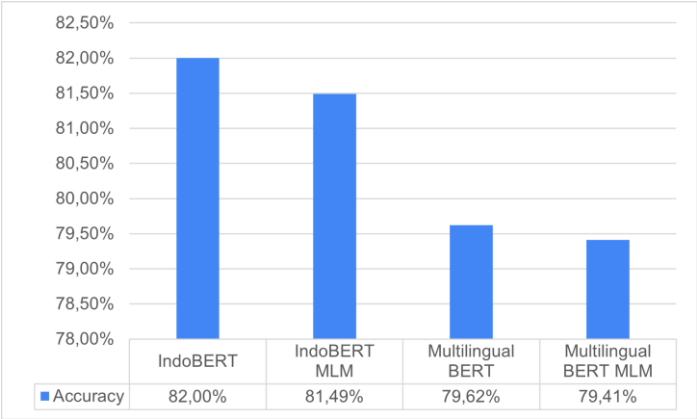

Figure 5. Average performance results of each model

Figure 5 illustrates the comparative accuracy of the four transformer-based models. IndoBERT achieved the highest classification performance with an average accuracy of 82%. This result aligns with previous studies that reported IndoBERT's superiority in handling Indonesian texts because the model is trained exclusively on large-scale Indonesian corpora (Pradnyana et al., 2023). As a monolingual model, IndoBERT can capture linguistic nuances, morphology, and syntactic structures

more effectively than multilingual models whose capacity must be divided among many languages. This linguistic specialization explains why IndoBERT consistently outperformed the other variants in this study.

IndoBERT with additional Masked Language Modeling (MLM) pre-training also produced competitive results with an accuracy of 81.49%. However, its performance was slightly lower than that of IndoBERT without MLM. This can be attributed to the limited volume of unlabeled disaster-related corpora used in the MLM stage, which was insufficient to significantly enhance the model's contextual understanding (Wei et al., 2024). Previous literature has emphasized that MLM pre-training generally improves performance (Joshi et al., 2020), yet this effect is strongly dependent on the size and quality of the corpus. In this case, the limited dataset constrained the potential gains from MLM.

Meanwhile, Multilingual BERT achieved a lower accuracy of 79.62%. This result is consistent with earlier studies that showed multilingual models often underperform compared to monolingual counterparts in language-specific tasks. The reason lies in Multilingual BERT's architecture, which is jointly trained on 104 languages, reducing Indonesian's representation depth. Consequently, its ability to capture fine-grained linguistic features in Indonesian is weaker compared to IndoBERT (Sebastian et al., 2022).

The lowest performance was recorded by Multilingual BERT with MLM pre-training, with an average accuracy of 79.41%. Similar to IndoBERT MLM, the limited size of the MLM training corpus restricted the potential improvement. Furthermore, because Multilingual BERT already allocates its representational capacity across more than one hundred languages, the additional MLM step did not provide substantial benefit and even introduced slight fluctuations in performance (Patel et al., 2022).

From the perspective of previous research, most comparative studies only examined IndoBERT and mBERT without involving MLM, such as Mahardika et al. (2023) who reported IndoBERT's superiority in hate speech classification with 88% accuracy compared to Multilingual BERT's 77%. To our knowledge, explicit comparisons between MLM and non-MLM variants for Indonesian-language classification tasks remain limited. This study, therefore, provides one of the first empirical analyses on the effect of MLM pre-training in IndoBERT and Multilingual BERT, offering new insights into the strengths and limitations of MLM in low-resource, disaster-related datasets.

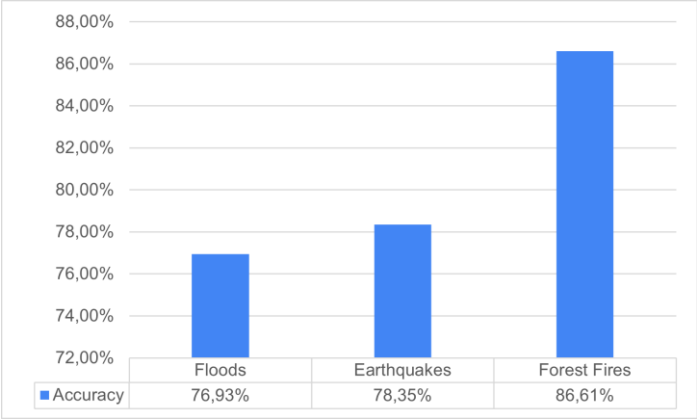

Figure 6. Average performance results of each dataset

Figure 6 shows that the highest result was obtained on the forest fire dataset, with an accuracy of 86.61%. This result may be due to the characteristics of the data being more precise or more explicit

in conveying disaster information, for example, the use of keywords in Indonesian, such as "asap" (smoke), "kebakaran" (fire), "api" (flame), or "hutan" (forest), which the model easily recognizes. On the other hand, the earthquake and flood datasets showed lower performance. This result may be due to greater ambiguity in the language used by users when conveying information about "gempa" (earthquake) or "banjir" (flood) on social media. For instance, the word "banjir" can be used in metaphorical or ironic contexts in Indonesian, such as "banjir promo" (flood of promotions) or "banjir tugas" (flood of assignments), which can make it difficult for the model to recognize the actual disaster context.

Several misclassifications were identified in the evaluation process, indicating that the models still encountered difficulties distinguishing between certain categories. For instance, the message “pemberitahuan libur dikarenakan banjir” was originally labeled as eyewitness, but the model incorrectly classified it as non-eyewitness. Similarly, the tweet “kyk ada gempa barusan” with the label don’t know was misclassified as non-eyewitness, showing that the model struggled to differentiate between uncertain expressions and informative content. Another example occurred in the forest fire dataset, where the message “asap kebakaran hutan di jambi semakin tebal” with the actual label non-eyewitness was predicted as eyewitness, suggesting that the model tended to focus on the descriptive mention of the disaster event rather than the source perspective.

These cases demonstrate that misclassification often arises due to the ambiguity of informal language, overlapping expressions across categories, and the limited ability of models to capture subtle contextual cues in short texts. Such findings emphasize the importance of refining preprocessing strategies and exploring more context-aware language models to improve classification accuracy in disaster-related messages.

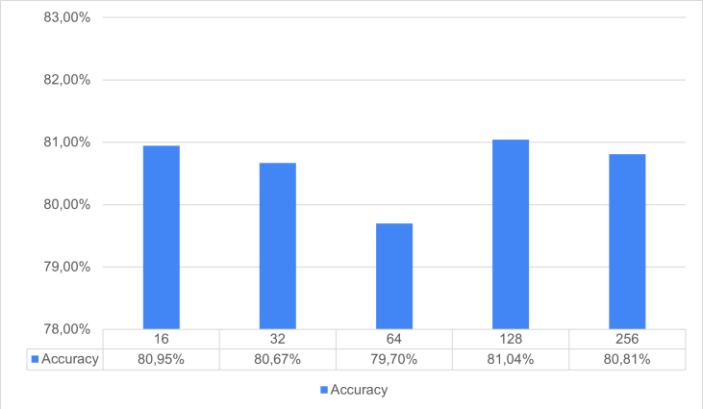

Figure 7. Average performance results for each batch size

Figure 7 presents an experiment with variations in batch size, showing that changes in batch size do not cause significant fluctuations in model performance. Almost all configurations tend to produce consistent accuracy, around 81% when rounded. This result can be explained by the fact that the BERT model is already relatively stable and able to adapt well after the pretraining process, so it is not significantly affected by changes in batch size within the range of 16 to 256. In larger batches, the representation of the three labels (eyewitness, non-eyewitness, and don't know) tends to be more balanced, which allows the learning process to be more stable and not too biased toward the majority class. However, this improvement is not drastic because a batch size that is too large can also reduce the model's ability to recognize patterns in general. In contrast, small batches can slow the training process to reach optimal results.

As a comparison, we compared our model performance results with previous research using the same dataset to determine the effectiveness of the Indonesian BERT-based model approach in improving the classification performance of disaster messages, as shown in Table 7. The best accuracy results from our study for each dataset surpassed the classification results using CNN and SVM models. These results may indicate that using IndoBERT and IndoBERT MLM provides superior performance in understanding and grouping Indonesian-language texts, thereby improving model performance.

Table 7. Comparison with previous research

Dataset	*Previous research*		*Our research*	
	Method	*Accuracy (%)*	*Method*	*Accuracy (%)*
Floods	2D CNN (Faisal et al., 2022)	78.33	IndoBERT	**80.67**
	SVM (Delimayanti et al., 2020)	77.87		
Earthquakes	2D CNN (Faisal et al., 2022)	73.88		**81.50**
	SVM (Nooralifa et al., 2021)	64.43		
Forest fires	2D CNN (Faisal et al., 2022)	81.97	IndoBERT MLM	**88.33**
	CNN (Rinaldi et al., 2021)	81.97		

From a technical perspective, Masked Language Modeling pre-training has not shown consistent performance improvements. Although theoretically, MLM can enrich the understanding of language context, this study's results show that MLM models sometimes produce lower accuracy than models without additional pre-training. In addition, the MLM pre-training process requires large computational resources, so its effectiveness still needs further evaluation for specific cases such as disaster message classification. Another limitation of this study is that the parameter exploration is limited, namely, only to batch size variation. Testing has not yet been conducted on other important parameters such as learning rate, number of epochs, or max length.

CONCLUSIONS AND RECOMMENDATIONS

The results of this study show that the IndoBERT model achieved the highest performance among others, with an average accuracy of 82%. The use of MLM on IndoBERT achieved an average accuracy of 81.49%. Meanwhile, the Multilingual BERT model achieved an average accuracy of 79.62%. Pre-trained MLM was also applied to Multilingual BERT, which reached an average accuracy of 79.41%. This result shows that adding pre-trained MLM does not always perform better than the model without additional pre-training. In addition, this study indicates that monolingual models such as IndoBERT can outperform multilingual models such as Multilingual BERT in classifying natural disaster messages on social media.

This study also found the best batch size for each natural disaster message dataset, such as in the flood dataset, where the best batch size is 16 or 256 using IndoBERT, achieving an accuracy of 80.67%, sensitivity of 80.67%, and specificity of 90.33%. In the earthquake dataset, the best batch size is 256 using IndoBERT, resulting in an accuracy of 81.50%, sensitivity of 81.50%, and specificity of 90.75%. Meanwhile, the best batch size for the forest fire dataset is 16 using IndoBERT MLM, achieving an accuracy of 88.33%, sensitivity of 88.33%, and specificity of 94.17%.

This study has limitations in the scope of parameter exploration, which only focused on batch size variation and the suboptimal impact of pre-training MLM on improving model performance. To address this, future research should apply hyperparameter tuning techniques to the pre-trained MLM model,

IndoBERT, Multilingual BERT, and several other BERT variants in the Indonesian language to improve model performance in classifying disaster messages on social media.

ACKNOWLEDGEMENT

This research was supported by funding from the DRTPM Research Program of Indonesia's Ministry of Education, Culture, Research, and Technology. Main Contract Number: 056/E5/PG.02.00.PL/2024, Derivative Contract Number: 1026/UN8.2/PG/2024.

REFERENCES

Amriza, R. N. S., Ngafidin, K. N. M., & Ratnasari, W. (2022). The Impact of Personal, Environmental, and Information Platform Factors on Disaster Information Sharing on Twitter. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *8*(2), 104–121. https://doi.org/10.26594/register.v8i2.2540

Aygun, I., Kaya, B., & Kaya, M. (2022). Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic With Deep Learning. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 2360–2369. https://doi.org/10.1109/JBHI.2021.3133103

Delimayanti, M. K., Sari, R., Laya, M., Faisal, M. R., Pahrul, & Naryanto, R. F. (2020). The Effect of Pre-Processing on the Classification of Twitter's Flood Disaster Messages Using Support Vector Machine Algorithm. *2020 3rd International Conference on Applied Engineering (ICAE)*, 1–6. https://doi.org/10.1109/ICAE50557.2020.9350387

Faisal, M. R., Budiman, I., Abadi, F., Haekal, M., Delimayanti, M. K., & Nugrahadi, D. T. (2022). Using Social Media Data to Monitor Natural Disaster: A Multi Dimension Convolutional Neural Network Approach with Word Embedding. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(6), 1037–1046. https://doi.org/10.29207/resti.v6i6.4525

Faisal, M. R., Budiman, I., Abadi, F., Nugrahadi, D. T., Haekal, M., & Sutedja, I. (2022). Applying Features Based on Word Embedding Techniques to 1D CNN for Natural Disaster Messages Classification. *2022 5th International Conference on Computer and Informatics Engineering, IC2IE 2022*, 192–197. https://doi.org/10.1109/IC2IE56416.2022.9970188

Faisal, M. R., Fitriani, K. E., Mazdadi, M. I., Indriani, F., Turianto Nugrahadi, D., & Prastya, S. E. (2025). Enhancing Natural Disaster Monitoring: A Deep Learning Approach to Social Media Analysis Using Indonesian BERT Variants. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(1), 77–89. https://doi.org/10.35882/ijeeemi.v7i1.38

Fuady, M., Munadi, R., & Fuady, M. A. K. (2021). Disaster mitigation in Indonesia: between plans and reality. *IOP Conference Series: Materials Science and Engineering*, *1087*(1), 012011. https://doi.org/10.1088/1757-899x/1087/1/012011

Garrido-Merchan, E. C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. (2023). Comparing BERT Against Traditional Machine Learning Models in Text Classification. *Journal of Computational and Cognitive Engineering*, *2*(4), 352–356. https://doi.org/10.47852/bonviewJCCE3202838

Gasmi, K. (2022). Improving Bert-Based Model for Medical Text Classification with an Optimization Algorithm. In J. and B. D. and H. B. and K. M. Bădică Costin and Treur (Ed.), *Advances in Computational Collective Intelligence* (pp. 101–111). Springer International Publishing.

Guo, Y., Xie, Z., Chen, X., Chen, H., Wang, L., Du, H., Wei, S., Zhao, Y., Li, Q., & Wu, G. (2022). ESIE-BERT: Enriching Sub-words Information Explicitly with BERT for Joint Intent Classification and SlotFilling. *ArXiv*. http://arxiv.org/abs/2211.14829

Ingkafi, D. A., Aryana, G. A., Putra, A. K., & Kusumaningrum, R. (2023). Sentiment Analysis of The National Covid-19 Vaccination Program on Twitter Using The Bidirectional Encoder Representation from Transformer. *ICIC Express Letters*, *17*(2), 201–208. https://doi.org/10.24507/icicel.17.02.201

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., Levy, O., & Allen, †. (2020). *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. https://doi.org/10.1162/tacl

Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-09381-9

Komara, D. A., & Hadiapurwa, A. (2022). Automating Twitter Data Collection: A Rapidminer-Based Crawling Solution. *Publis Journal Publication Library and Information Science*, *6*.

Koroteev MV. (2021). *BERT: A Review of Applications in Natural Language Processing and Understanding*. https://doi.org/10.48550/arXiv.2103.11943

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. https://doi.org/10.48550/arXiv.2011.00677

Kumar, K. A., & Renuka, G. A. (2024). Leveraging Bidirectional Encoder Representations from Transformers (BERT) for Enhanced Sentiment Analysis. In S. and R. S. S. Chillarige Raghavendra

Rao and Distefano (Ed.), *Advances in Computational Intelligence and Informatics* (pp. 87–95). Springer Nature Singapore.

Kumar, S. (2024). Text Normalization. In *Python for Accounting and Finance* (pp. 133–145). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54680-8_9

Ma, E. S. (2023). Investigating Masking-based Data Generation in Language Models. *ArXiv*. http://arxiv.org/abs/2307.00008

Mahardika, M. R., Wijaya, I. P. J., Prayoga, A. R., Lucky, H., & Iswanto, I. A. (2023). Exploring the Performance of BERT Models for Multi-Label Hate Speech Detection on Indonesian Twitter. *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 256–261. https://doi.org/10.1109/AiDAS60501.2023.10284596

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2022). BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science*, *216*, 714–721. https://doi.org/10.1016/j.procs.2022.12.188

Nabiilah, G. Z., & Suhartono, D. (2023). Personality Classification Based on Textual Data using Indonesian Pre-Trained Language Model and Ensemble Majority Voting. *Revue d'Intelligence Artificielle*, *37*(1), 73–81. https://doi.org/10.18280/ria.370110

Noor Fakhruzzaman, M., Zahrotul Jannah, idah, Ardiati Ningrum, R., & Fahmiyah, I. (2021). *Clickbait Headline Detection in Indonesian News Sites using Multilingual Bidirectional Encoder Representations from Transformers (M-BERT)*. https://doi.org/10.48550/arXiv.2102.01497

Nooralifa, S. M., Faisal, M. R., Muliadi, M., Abadi, F., & Nugroho, R. A. (2021). Identifikasi Otomatis Pesan Saksi Mata pada Media Sosial Saat Bencana Gempa. *KLIK KUMPULAN J. ILMU Komput [Komputer Klik-Compilation]*, *8*, 129–138. https://doi.org/10.20527/klik.v8i2.351

Patel, A., Oza, P., & Agrawal, S. (2022). Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. *Procedia Computer Science*, *218*, 2459–2467. https://doi.org/10.1016/j.procs.2023.01.221

Pradnyana, G. A., Anggraeni, W., Yuniarno, E. M., & Purnomo, M. H. (2023). Fine-Tuning IndoBERT Model for Big Five Personality Prediction from Indonesian Social Media. *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 93–98. https://doi.org/10.1109/ISITIA59021.2023.10221074

Pramana, R., Jonathan, M., Yani, H. S., & Sutoyo, R. (2024). A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews. *Procedia Computer Science*, *245*(C), 399–408. https://doi.org/10.1016/j.procs.2024.10.266

Rahman Isnain, A., Hendrastuty, N., & Andraini, L. (2021). Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis. *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, *6*(1). https://doi.org/10.30591/jpit.v6i1.3245

Rinaldi, Faisal, M. R., Mazdadi, M. I., Nugroho, R. A., & Abadi, F. (2021). Eye Witness Message Identification on Forest Fires Disaster Using Convolutional Neural Network. *Journal of Data Science and Software Engineering*, *2*. http://fb.me/6sFXlyEcj

Sadaiyandi, J., Arumugam, P., Sangaiah, A. K., & Zhang, C. (2023). Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset. *Electronics (Switzerland)*, *12*(21). https://doi.org/10.3390/electronics12214423

Sebastian, D., Purnomo, H. D., & Sembiring, I. (2022). BERT for Natural Language Processing in Bahasa Indonesia. *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 204–209. https://doi.org/10.1109/ICICyTA57421.2022.10038230

Shidik, G. F., Saputra, F. O., Saraswati, G. W., Winarsih, N. A. S., Rohman, M. S., Pramunendar, R. A., Kusuma, E. J., Ratmana, D. O., Venus, V., Andono, P. N., & Hasibuan, Z. A. (2024). Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM). *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(3). https://doi.org/10.1016/j.joitmc.2024.100358

Tanvir, H., Kittask, C., Eiche, S., & Sirts, K. (2021). EstBERT: A Pretrained Language-Specific BERT for Estonian. *ArXiv*. https://doi.org/10.48550/arXiv.2011.04784

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Uliniansyah, M. T., Budi, I., Nurfadhilah, E., Afra, D. I. N., Santosa, A., Latief, A. D., Jarin, A., Gunarso, Jiwanggi, M. A., Hidayati, N. N., Fajri, R., Suryono, R. R., Pebiana, S., Shaleha, S., Ramdhani, T. W., & Sampurno, T. (2024). Twitter dataset on public sentiments towards biodiversity policy in Indonesia. *Data in Brief*, *52*. https://doi.org/10.1016/j.dib.2023.109890

Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment Analysis of Twitter Data. In *Applied Sciences (Switzerland)* (Vol. 12, Issue 22). MDPI. https://doi.org/10.3390/app122211775

Wei, C., Wang, Y.-C., Wang, B., & Kuo, C.-C. J. (2024). An Overview of Language Models: Recent Developments and Outlook. *APSIPA Transactions on Signal and Information Processing*, *13*(2). https://doi.org/10.1561/116.00000010

Wettig, A., Gao, T., Zhong, Z., & Chen, D. (2022). Should You Mask 15% in Masked Language Modeling? *ArXiv*. http://arxiv.org/abs/2202.08005

Wu, S. (2022). How Do Multilingual Encoders Learn Cross-lingual Representation? *ArXiv*. https://doi.org/10.48550/arXiv.2207.05737

Yang, S., & Yang, Q. (2025). Joint pairwise learning and masked language models for neural machine translation of English. *Artificial Life and Robotics*. https://doi.org/10.1007/s10015-025-01008-2