Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

Classification of Eyewitness Social Media Messages for Natural Disaster Monitoring using BERT Variants

Muhammad Bashir Hanafi[1], Mohammad Reza Faisal*[2], Friska Abadi[3], Irwan Budiman[4], Setyo Wahyu Saputro[5], Njideka Nkemdilim Mbeledogu[6]

[1,2,3,4,5]Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Indonesia
[6]Department of Computer Science, Nnamdi Azikiwe University, Anambra State, Nigeria

Email: [2]reza.faisal@ulm.ac.id

Received : Sep 22, 2025; Revised : Jan 5, 2026; Accepted : Jan 19, 2026; Published : Jun 15, 2026

Abstract

The rapid growth of disaster-related social media data demands effective monitoring. However, its real-time source presents challenges due to large volumes of unstructured and noisy data. This study aims to improve effective monitoring with BERT variants to classify eyewitness reports on Twitter/X. Earlier studies have applied machine-learning and deep-learning models to automate the monitoring of eyewitness messages on social media, but these models still have shortcomings. Traditional machine-learning models rely on handcrafted and frequency-based features, limiting their ability to capture contextual semantics. Deep-learning models offer improved performance but still face challenges in modeling long-range dependencies and handling high-volume social media streams. This issue is pronounced in social media streams. This study employs transformer-based models using several BERT variants (BERT, RoBERTa, DistilBERT, ELECTRA, and ALBERT). Each model is pre-trained with the Masked Language Modeling (MLM) objective, and batch-size optimization is applied to boost performance. Experimental results indicate that a batch size of 16 consistently yields the best performance, with the standard BERT model achieving the highest macro-F1 score of 0.762. By disaster type, macro-F1 scores reach 0.744 for hurricane, 0.793 for flood, 0.756 for earthquake, and 0.750 for wildfire. BERT (16) outperforms the other BERT variants and twelve baseline models from prior research. Unlike previous approaches, this study leverages pre-trained Masked Language Models to optimize classification on disaster-related datasets. The findings contribute to the development of transformer-based architectures for text classification in real-time disaster informatics, leading to more accurate situational awareness and reduced delays in emergency decision-making.

Keywords : *BERT, Classification, Masked Language Modeling, Transformer, Twitter/X*

1. INTRODUCTION

Natural disasters are a significant global challenge that is worsened by climate change, rapid urbanisation, and environmental degradation. The Emergency Event Database (EM-DAT) recorded 393 natural disasters in 2024, impacting more than 167.2 million people in various parts of the world [1]. The increased frequency and intensity of disasters require a more effective, fast, and scalable response to reduce the impact of casualties and material losses. In this effort, the use of modern technology, such as communication systems [2], Internet of Things (IoT) [3], and data analysis [4] has shown significant results in overcoming various challenges posed by natural disasters. Social media, a form of technology used as a communication tool, can obtain information in real-time by taking advantage of the massive volume of data users upload during natural disasters. For example, users shared reports on social media during a major flood disaster in southern Brazil to prompt faster responses from authorities and humanitarian organizations [5]. However, handling the data from social media is still a big challenge because of the large volume, unstructured nature, and tendency to be noisy, such as new, non-standard, and ambiguous words [6]. In

particular, within the context of natural disasters, one of the key difficulties is identifying eyewitness reports.

Various strategies can be applied to address these challenges, and one widely explored direction is the use of machine learning and deep learning methods. Previous studies have used machine learning [7], [8], [9], [10] and deep learning [11] methods using a natural disaster dataset [7] to identify eyewitness reports and have shown improved performance. However, machine learning methods still use manual and frequency-based feature extraction, which makes it challenging to capture semantic similarity and ignore the ambiguity that ignores the word sequence in a sentence [12]. Deep learning architectures that have been explored, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), can improve performance but still face difficulties handling social media texts due to the inability to handle Out-of-Vocabulary (OOV) words. CNN tends to be limited to short texts and has difficulty capturing word sequences [13] LSTMs are better for long texts, but still face problems with very long-term dependencies. Moreover, LSTMs still operate sequentially, which makes the computation process slow, as the processing is still done word by word in a sentence [14]. This gap highlights the need for a new approach to handle the massive volume of social media, unstructured text, and noise, such as newly coined, non-standard, and ambiguous words.

Natural Language Processing (NLP) developments continue to progress with the advent of transformer architecture-based models, which have been essential milestones in developing NLP. The transformer adopts an architecture that allows for parallel sentence processing, in contrast to the sequential processing approach applied by previous models. In addition, the transformer utilizes a self-attention mechanism, which allows the model to consider every word in a sentence simultaneously, thereby improving the model's ability to capture semantic context more effectively [15]**.** Bidirectional Encoder Representations from Transformers (BERT) has further advanced the capabilities of transformer-based models and has provided significant advances with excellent performance compared to other models [16], [17]. BERT uses a bidirectional self-attention mechanism that enables deep understanding of contextual relationships in the text [18]. During the pre-training phase, the Masked Language Modeling (MLM) strategy is applied to understand the word relationship and predict hidden tokens in sentences to capture contextual semantics deeply [19], [20].

BERT has model variants, such as RoBERTa, DistilBERT, ELECTRA, and ALBERT, which are state-of-the-art in developing transformer-based models to overcome the limitations of the original BERT model. RoBERTa, for example, optimises BERT with pre-trained dynamic masking, eliminating Next Sentence Prediction (NSP), larger batch size, and using Byte Pair Encoding (BPE) [21], [22]. Meanwhile, DistilBERT was developed in a smaller size through the knowledge-distillation process [23]. ELECTRA uses a different pre-training approach using the Generator-Discriminator to replace tokens [24]. ALBERT reduces the number of parameters without sacrificing performance [25]. Although various advances in transformer models have been achieved, research on the BERT variants model is still relatively limited for identifying eyewitness reports on social media during disasters. Implementing MLM to the BERT variants can also open up opportunities, enabling the model to deeply understand the relationship between words in the specific dataset [7].

This study aims to identify the optimal BERT variant and batch configuration to enhance the classification of eyewitness messages. Unlike earlier works that focused only on machine learning and deep learning models, this research emphasizes how model optimization, such as batch size and pre-training strategy, affects the performance of social media disaster messages classification. The novelty of this research lies in implementing an MLM strategy with a natural disaster dataset for classifying eyewitness reports on social media. The research contributes in four aspects:

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

- Recommend the optimal batch size for eyewitness classification that can apply to all BERT variants and datasets for natural disasters.
- Evaluate the effect of MLM pre-trained models on the performance of BERT variants.
- Finding the best BERT variants model for classification.
- Contribute to the advancement of NLP development tailored to the classification of eyewitness identification.

2. METHOD

This section explains the stage of the methodology used, including research procedures using the BERT variant and using the BERT variant with Masked Language Model (MLM) model, as illustrated in Figure 1. To support the implementation of these models, the experiments were conducted in a computational environment with an Intel® Xeon® processor, an NVIDIA A100 Tensor Core GPU, and 80 GB of RAM.

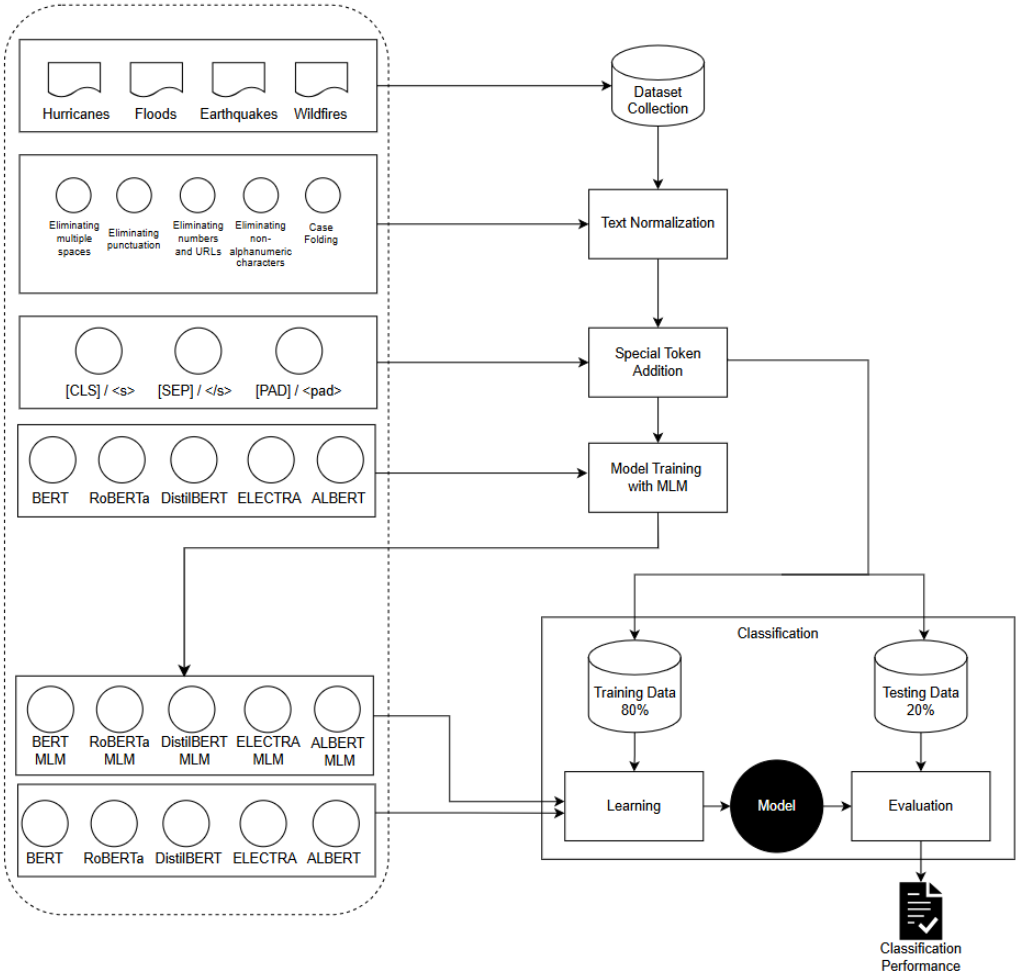


Figure 1. Research Procedure consisting of data collection, text normalization, special token addition, model training with MLM, classification, and evaluation

2.1. Dataset Collection

This study uses the natural disasters dataset [7] from the social media Twitter (now called X) [21]. Data is collected through the Twitter Streaming API with manual analysis and crowdsourced methods. This

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

dataset was chosen because it covers a diverse time range and region, uses the universal language (English), and reflects the class imbalance common in natural disaster events on social media. Each natural disaster dataset has three classes: direct_eyewitnesses, noneyewitnesses, and don't_know. Details of the dataset can be seen in Table 1.

Table 1. Natural Disasters Dataset Detail [7]

Dataset	Total	direct_eyewitnesses	noneyewitnesses	don't_know
Hurricanes	2000	465	1199	336
Floods	2000	627	551	822
Earthquakes	2000	1600	200	200
Wildfires	2000	189	1379	432

2.2. Text Normalization

Text normalization is performed to clean and convert text from an inconsistent format into a standard form. Natural disaster datasets are processed using the stages commonly used in text classification [11]. This process includes four steps: (1) eliminating multiple spaces, (2) eliminating punctuation, (3) eliminating numbers and URLs, (4) eliminating non-alphanumeric characters, and (5) case folding. These steps ensure that the model can process the text without irrelevant elements.

2.3. Special Token Addition

After text normalization, the following step is to add special tokens. This process involves including the [CLS] or <s> token to specify the beginning of the input text, the [SEP] or </s> token as a separator within the text, and the [PAD] or <pad> token to ensure all sequences have a consistent length within the input batch.

2.4. Model Training with MLM

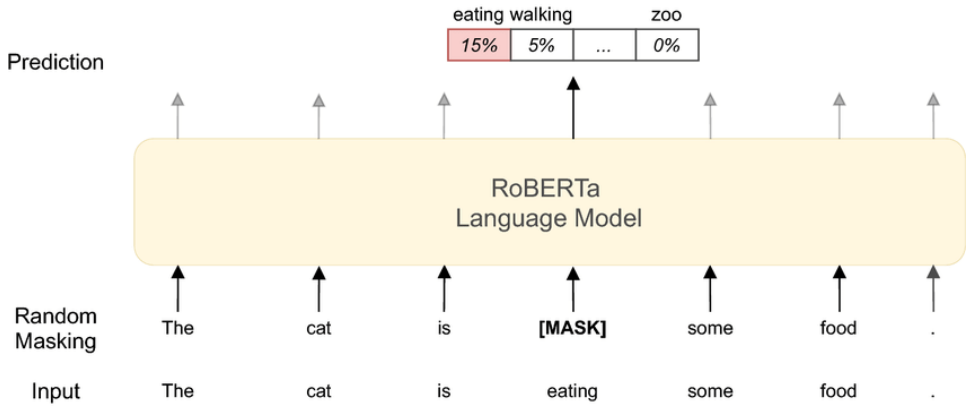


Figure 2. Masked Language Modeling (MLM) RoBERTa Architecture [21]

In Figure 2, the model is trained using Masked Language Modeling (MLM) with the unlabeled text natural disasters corpus [7] to understand the relationship between words in a specific dataset. The trained model can be used for classification tasks. Figure 2 shows an example of an architecture of MLM on RoBERTa, using the hiding of some tokens with [MASK] tokens. Next, the model was trained to predict hidden tokens by considering the context. The parameter configuration for MLM training includes batch

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

size 16, epoch 10, learning rate 2e-5 with AdamW optimizer, max length 128, and MLM probability 15% [19].

2.5. Classification

In the next step, each dataset will be divided into training and testing with an 80:20 ratio [26]. Training data is used to train the model, while testing data is used to measure model performance. The classification models to be used are divided into two types: the pre-trained BERT variant model: (1) BERT, (2) RoBERTa, (3) DistilBERT, (4) ELECTRA, (5) ALBERT, and the pre-trained MLM variant model: (6) BERT MLM, (7) RoBERTa MLM, (8) DistilBERT MLM, (9) ELECTRA MLM, (10) ALBERT MLM. These BERT variants offer different architectural strengths that are expected to improve addressing challenges such as high variability, non-standard texts, and noisy contexts. BERT serves as a baseline with strong bidirectional contextual representation. RoBERTa, with its dynamic masking and larger training corpus, is expected to provide more robust contextual representational that help interpret non-standard and ambiguous expressions commonly found in disaster-related posts. DistilBERT, with its reduced size, aims to retain contextual understanding while offering faster inference. ELECTRA, through a generator-discriminator pre-training objective, is expected to perform well with limited or noisy data by efficiently learning token replacements, and ALBERT reduced parameters through factorised embeddings and parameter sharing, enabling efficient training while maintaining strong performance, which is beneficial for handling large-scale social media streams. With an additional Masked Language Modelling pre-training objective, these models can be expected to improve the performance.

The parameter configuration for the model classification task includes AdamW optimizer, learning batch size {16, 32, 64, 128, 256}, epochs 10, learning rate 2e-5 with AdamW optimizer, and max length 128. A wide range of batch sizes was used to evaluate the effect of batch size. The number of epochs was set to 10, followed by finding [27], which demonstrated that various BERT-based models typically achieve optimal performance around this point. The learning rate of 2e-5, which is one of the default values for the BERT model, is a phenomenon in which the model forgets previously learned knowledge when trained on new data [28]. The maximum input length is set to 128 based on observations that the maximum input length falls below this threshold, allowing faster training without sacrificing information. This study employed a straightforward fixed 80:20 train-test split, and each model was assessed independently to ensure consistent comparison across all architectures with the limited computation.

2.6. Evaluation

After training, models will be evaluated on the testing data to measure the performance in classifying eyewitnesses of natural disasters in the four datasets, such as hurricanes, floods, earthquakes, and wildfires, into three classes: direct_eyewitnesses, non-eyewitnesses, and don't_know. The evaluation uses a confusion matrix to calculate the F1 score, which is used as a metric to determine the model's ability to handle class imbalances in the natural disasters dataset. In contrast, the macro F1 score is the average F1 score for all classes, regardless of the count of samples in each class, avoiding bias towards the majority class [29]. The overall macro F1 score is used to evaluate the model's performance across all datasets in the classification task [30]. The formula can be seen in (1)-(3).

$$F1\ score = 2\frac{precision\ .\ recall}{precision+recall} \quad (1)$$

$$macro\ F1\ score = \frac{1}{n}\sum_{i=1}^{n} F1\ score \quad (2)$$

$$macro\ F1\ score_{overall} = \frac{1}{n}\sum_{i=1}^{n} macro\ F1\ score \quad (3)$$

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

3. RESULT

3.1. Data Collection

The experiments in this study were conducted using the natural disaster dataset [7]. The dataset contains four disaster categories and three class labels, with the full distribution presented in Table 1. Examples of flood-related messages in the dataset are presented in Table 2.

Table 2. Examples of Flood-related Messages

Text	Label
Ex cyclone Debbie is bringing some bad weather where I am and so all schools in SE QLD are closed. So I have early holidays. YAY!!! 3	direct_eyewitnesses
Videos show Cyclone Debbie lashing north Queensland towns: https://t.co/sUkgAIU9Gj - Just In #Latest	noneyewitnesses
Oh #fudgethelab! Way to go, Swift Water team/emergency response teams :). https://t.co/LZO25yDqMe	don't-know

3.2. Text Normalization

After the data was collected, several standard preprocessing steps were applied. These steps follow common text-classification preprocessing procedures [11], including eliminating multiple spaces, eliminating punctuation, eliminating numbers and URLs, eliminating non-alphanumeric characters, and case folding. Examples of raw and normalized text are shown in Table 3.

Table 3. Examples of Raw and Normalized Text

Text	Text Normalization
Ex cyclone Debbie is bringing some bad weather where I am and so all schools in SE QLD are closed. So I have early holidays. YAY!!! 3	ex cyclone debbie is bringing some bad weather where i am and so all schools in se qld are closed so i have early holidays yay
Videos show Cyclone Debbie lashing north Queensland towns: https://t.co/sUkgAIU9Gj - Just In #Latest	videos show cyclone debbie lashing north queensland towns just in latest
Oh #fudgethelab! Way to go, Swift Water team/emergency response teams :). https://t.co/LZO25yDqMe	oh fudgethelab way to go swift water teamemergency response teams

3.3. Special Token Addition

Table 4 illustrates the procedure for special token addition in pre-trained BERT, BERT-MLM, and pre-trained RoBERTa, which the model will use. Pre-trained BERT and BERT-MLM produce identical special-token formatting due both models share the same WordPiece tokenizer and input representations. In contrast, pre-trained RoBERTa applies different special tokens due to architectural modifications and the use of a byte-level BPE tokenizer, which leads to different token segmentation patterns.

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

Table 4. Special Token Addition

Model	Text	Adding Special Token
BERT	ex cyclone debbie is bringing some bad weather where i am and so all schools in se qld are closed so i have early holidays yay	'[CLS]', 'ex', 'cyclone', 'debbie', 'is', 'bringing', 'some', 'bad', 'weather', 'where', 'i', 'am', 'and', 'so', 'all', 'schools', 'in', 'se', 'q', '##ld', 'are', 'closed', 'so', 'i', 'have', 'early', 'holidays', 'ya', '##y', 'a', '[SEP]', '[PAD]' … '[PAD]'
BERT MLM	ex cyclone debbie is bringing some bad weather where i am and so all schools in se qld are closed so i have early holidays yay	'[CLS]', 'ex', 'cyclone', 'debbie', 'is', 'bringing', 'some', 'bad', 'weather', 'where', 'i', 'am', 'and', 'so', 'all', 'schools', 'in', 'se', 'q', '##ld', 'are', 'closed', 'so', 'i', 'have', 'early', 'holidays', 'ya', '##y', 'a', '[SEP]', '[PAD]' … '[PAD]'
RoBERTa	ex cyclone debbie is bringing some bad weather where i am and so all schools in se qld are closed so i have early holidays yay	'<s>', 'ex', 'Ä cycl', 'one', 'Ä deb', 'bie', 'Ä is', 'Ä bringing', 'Ä some', 'Ä bad', 'Ä weather', 'Ä where', 'Ä i', 'Ä am', 'Ä and', 'Ä so', 'Ä all', 'Ä schools', 'Ä in', 'Ä se', 'Ä q', 'ld', 'Ä are', 'Ä closed', 'Ä so', 'Ä i', 'Ä have', 'Ä early', 'Ä holidays', 'Ä y', 'ay', 'Ä a', '</s>', '<pad>', …, '<pad>'

The tokenization process begins by inserting several tokens, which include the insertion of special tokens [CLS] or <s> tokens at the start of sentences, [SEP] or </s> tokens at the end of sentences, and [PAD] or <pad> tokens to ensure consistent input length within a batch.

3.4. Model Training with MLM

Table 5. Masking Process

Before Masking	After Masking
'[CLS]', 'ex', 'cyclone', 'debbie', 'is', 'bringing', 'some', 'bad', 'weather', 'where', 'i', 'am', 'and', 'so', 'all', 'schools', 'in', 'se', 'q', '##ld', 'are', 'closed', 'so', 'i', 'have', 'early', 'holidays', 'ya', '##y', 'a', '[SEP]', '[PAD]' … '[PAD]'	'[CLS]', '[MASK]', '[MASK]', 'debbie', 'is', 'bringing', 'some', '[MASK]', 'weather', 'where', 'i', 'am', 'and', 'so', '[MASK]', 'schools', 'in', 'se', 'q', '##id', 'are', 'closed', 'so', 'i', 'have', 'early', 'holidays', 'ya', '##y', 'a', '[SEP]', '[PAD]', ..., '[PAD]'
'[CLS]', 'videos', 'show', 'cyclone', 'debbie', 'lash', '##ing', 'north', 'queensland', 'towns', 'just', 'in', 'latest', '[SEP]', '[PAD]', ..., '[PAD]'	'[CLS]', 'videos', 'show', 'cyclone', [MASK], 'lash', '##ing', 'north', 'queensland', 'towns', 'just', 'in', 'latest', '[SEP]', '[PAD]', ..., '[PAD]'
'[CLS]', 'oh', 'fu', '##dget', '##hel', '##ab', 'way', 'to', 'go', 'swift', 'water', 'team', '##eme', '##rgen', '##cy', 'response', 'teams', '[SEP]', '[PAD]', ..., '[PAD]'	'[CLS]', 'oh', 'fu', '##dget', [MASK],[MASK], [MASK], 'to', 'go', 'swift', 'water', 'team', '##eme', '##rgen', '##cy', 'response', '[MASK], '[SEP]', '[PAD]', ..., '[PAD]'

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

After adding the special tokens, the model was trained using Masked Language Modeling (MLM), which enables the model to learn contextual relationships between words. In this process, some words in the sentence are randomly replaced with the [MASK] token, and the model is trained to predict the masked words. Table 5 shows the implementation of the masking process in the MLM model.

3.5. Classification

This sub-chapter describes the classification result of eyewitness identification in the natural disasters dataset. This study employs the F1 score for each class and the macro F1 score to account for class imbalance by averaging F1 scores across all classes equally. No explicit class balancing techniques, such as oversampling or synthetic data generation applied in this study to preserve the original data distribution, which reflects realistic disaster-related communication scenarios. Table 6 illustrates the result of the model's performance.

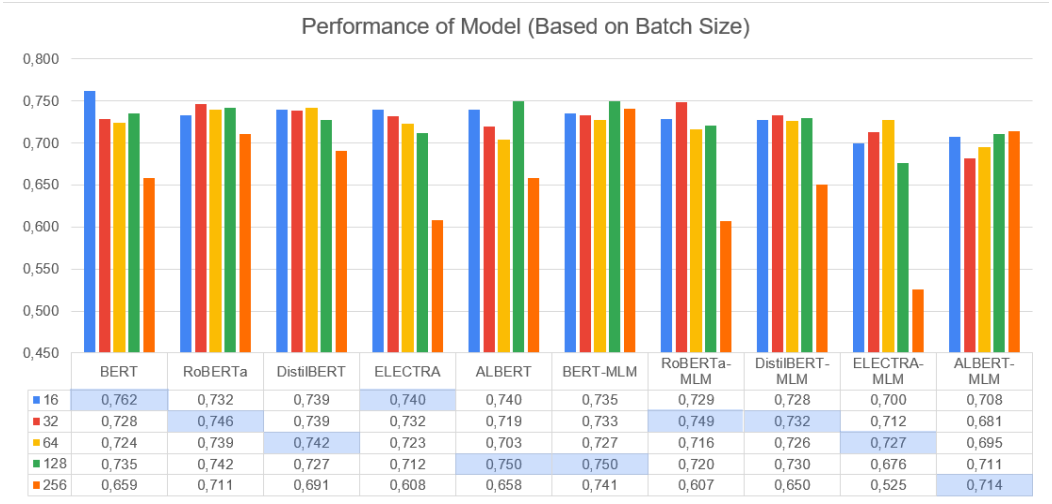
Table 6. Overall Results of The Classification Model (Macro F1 Score)

Model	Batch Size	Hurricanes	Floods	Earthquakes	Wildfires
BERT	16	0.744	0.793	0.759	0.751
	32	0.731	0.769	0.704	0.709
	64	0.708	0.758	0.688	0.742
	128	0.758	0.751	0.710	0.722
	256	0.730	0.772	0.549	0.584
RoBERTa	16	0.757	0.740	0.680	0.752
	32	0.740	0.762	0.724	0.758
	64	0.761	0.783	0.663	0.750
	128	0.752	0.776	0.703	0.738
	256	0.772	0.735	0.664	0.673
DistilBERT	16	0.734	0.756	0.714	0.754
	32	0.738	0.773	0.710	0.735
	64	0.736	0.772	0.710	0.751
	128	0.737	0.770	0.665	0.736
	256	0.731	0.761	0.587	0.685
ELECTRA	16	0.683	0.795	0.726	0.755
	32	0.705	0.744	0.737	0.742
	64	0.718	0.760	0.688	0.726
	128	0.725	0.724	0.700	0.699
	256	0.700	0.663	0.541	0.530
ALBERT	16	0.741	0.778	0.732	0.707
	32	0.726	0.784	0.648	0.719
	64	0.750	0.759	0.634	0.671
	128	0.739	0.774	0.764	0.724
	256	0.724	0.769	0.621	0.520
BERT MLM	16	0.761	0.761	0.686	0.733
	32	0.761	0.738	0.711	0.722
	64	0.760	0.750	0.711	0.685
	128	0.775	0.758	0.720	0.747
	256	0.743	0.748	0.713	0.759

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

Model	Batch Size	Hurricanes	Floods	Earthquakes	Wildfires
RoBERTa MLM	16	0.725	0.749	0.692	0.748
	32	0.745	0.756	0.733	0.762
	64	0.741	0.743	0.664	0.716
	128	0.729	0.747	0.677	0.728
	256	0.714	0.650	0.531	0.532
DistilBERT MLM	16	0.760	0.747	0.715	0.687
	32	0.765	0.757	0.690	0.717
	64	0.740	0.767	0.686	0.712
	128	0.752	0.772	0.704	0.692
	256	0.757	0.738	0.558	0.548
ELECTRA MLM	16	0.718	0.748	0.635	0.698
	32	0.683	0.770	0.681	0.716
	64	0.724	0.776	0.703	0.705
	128	0.701	0.742	0.674	0.590
	256	0.645	0.638	0.296	0.521
ALBERT MLM	16	0.705	0.739	0.663	0.724
	32	0.674	0.727	0.675	0.650
	64	0.701	0.717	0.694	0.669
	128	0.709	0.702	0.723	0.710
	256	0.701	0.713	0.718	0.725

Figure 3 shows the optimal configurations that yielded the best performance after applying the averaging to each batch size, providing a clear insight into the superior hyperparameters for each specific language model architecture.



	BERT	RoBERTa	DistilBERT	ELECTRA	ALBERT	BERT-MLM	RoBERTa-MLM	DistilBERT-MLM	ELECTRA-MLM	ALBERT-MLM
16	0,762	0,732	0,739	0,740	0,740	0,735	0,729	0,728	0,700	0,708
32	0,728	0,746	0,739	0,732	0,719	0,733	0,749	0,732	0,712	0,681
64	0,724	0,739	0,742	0,723	0,703	0,727	0,716	0,726	0,727	0,695
128	0,735	0,742	0,727	0,712	0,750	0,750	0,720	0,730	0,676	0,711
256	0,659	0,711	0,691	0,608	0,658	0,741	0,607	0,650	0,525	0,714

Figure 3. Optimal Configuration of Each Model

3.6. Evaluation

Figure 4 shows an example of a confusion matrix of the BERT with a batch size of 16 on the hurricane disaster dataset. In this study, the confusion matrix is used to calculate precision and recall, which are then combined to obtain the F1 score as the primary evaluation metric.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

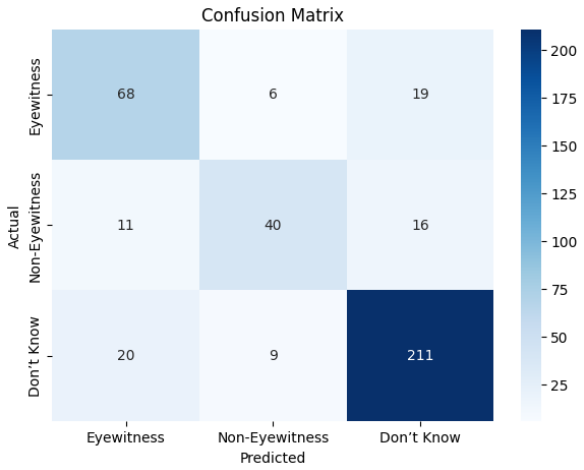


Figure 4. Confusion Matrix

Table 7 summarizes examples of the evaluation metrics for each class and presents the macro F1 score of the BERT model (batch size 16) on the hurricane dataset. The don't-know class obtained the highest F1 score, followed by direct_eyewitness and noneyewitnesses. Since the macro F1 score is calculated as the average F1 score across all classes, the performance of each class directly contributes to the final macro value.

Table 7. Summary of BERT (16) on the Hurricane Dataset

Class	TP	FP	FN	Precision	Recall	F1 Score
direct_eyewitnesses	68	31	25	0.687	0.731	0.708
noneyewitnesses	40	15	27	0.727	0.597	0.656
don't-know	211	35	29	0.858	0.879	0.869
Macro F1 Score						**0.744**

4. DISCUSSIONS

The optimal batch size configuration was determined by averaging the performance on each type of natural disaster and each model variant simultaneously, as in Figures 5 and 6. The analysis results show that 16 is the optimal batch size for producing the best average model performance. This finding indicates that a batch size of 16 performs well in classifying eyewitness messages related to natural disasters. Larger batch sizes (32, 64, 128) resulted in relatively lower average performance than smaller ones. Meanwhile, the average performance significantly declined when the batch size was too large (256). This is due to the batch configuration of 16 finds the optimal balance between training stability and regularization effects to adapt the model to varying social media texts.

The overall performance is shown in Figure 7 using an optimal batch size (16) from the hurricanes, floods, earthquakes, and wildfires dataset. It can be seen that each classification model has a different performance. In general, BERT variant models with MLM training showed that all BERT variants that did not use MLM training showed higher performance than all BERT variants that used MLM training. These findings suggest that MLM training does not always positively impact model performance.

Based on the paired two-tailed t-test resulted in a p-value of 0.204, indicating no statistically significant difference between MLM and original variants. It can be explained that the lower performance is due to the differences in the pre-training process adopted by each model.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

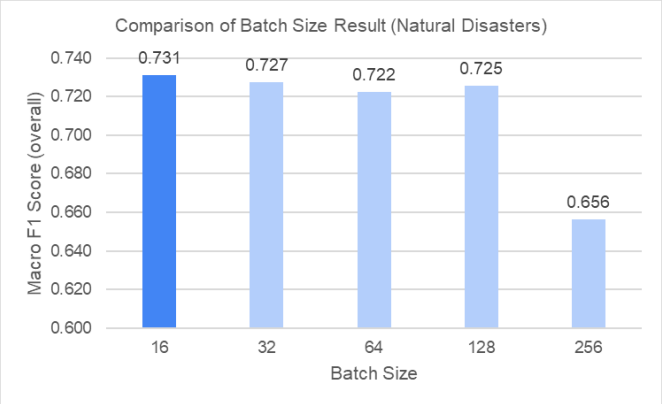


Figure 5. Comparison of Batch Size Result (Natural Disasters)

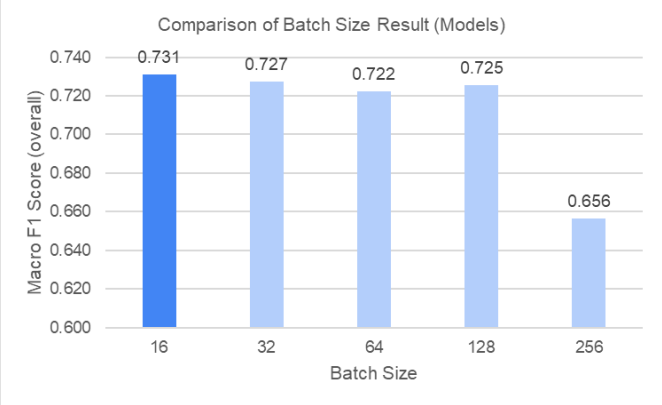


Figure 6. Comparison of Batch Size Result (Models)

For example, ALBERT has different pre-training techniques, namely using factorized embedding parameters and cross-layer parameter sharing. In contrast, ELECTRA uses replaced token detection as the core of the pre-training. This indicates that applying additional MLM-based pre-training across different architectures may result in representation mismatch or exacerbate overfitting during model adaptation.

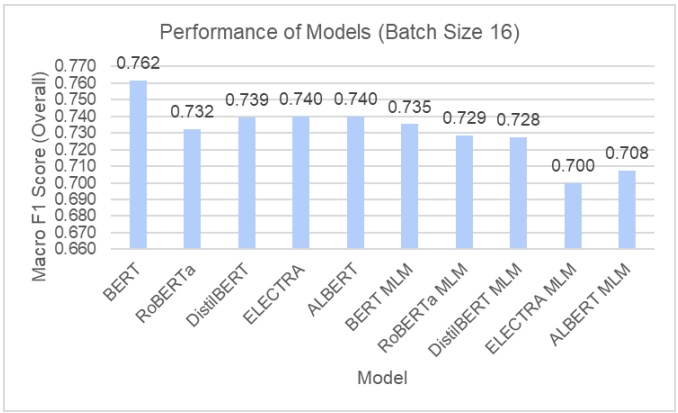


Figure 7. Performance of Models with Batch Size 16

Based on Figure 7, it can be seen that without pre-training MLM, BERT (16) produces the highest performance. In addition, ALBERT and ELECTRA achieved the same performance. However, RoBERTa has the lowest performance among the other BERT variant models. The advantages of the BERT (16) model

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

are based on its ability to process bidirectionally and pre-train Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

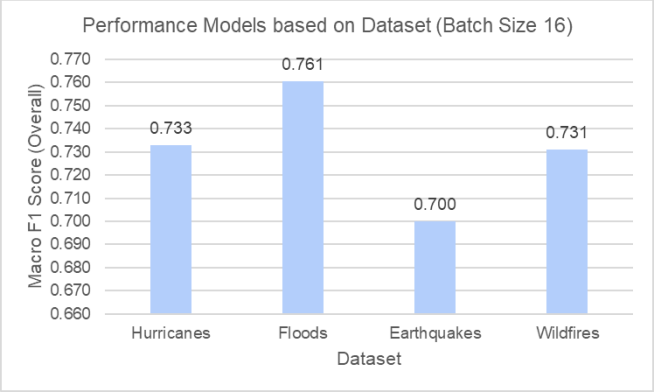


Figure 8. Performance Models based on Dataset with Batch Size 16

Meanwhile, other models have different performance due to differences in representation generated during the pre-training period, which causes differences in performance when fine-tuning specific datasets. In models with pre-training MLM, BERT MLM (16) produces the highest performance, followed by RoBERTa MLM and DistilBERT MLM. However, the ALBERT MLM and ELECTRA MLM models experienced significant declines compared to other models. This is because the BERT model was initially optimized using MLM training strategies, so retraining using MLM was in line with the pre-training strategy. BERT (16) generally shows the highest performance among other models. This indicates that the basic model of the BERT variant with precise hyperparameter adjustment can better capture semantic and contextual information.

Figure 8 shows the performance of the BERT variants based on the natural disasters categories. The BERT variants model performs best in the flood category, indicating that it is most effective in classifying flood-related datasets. Meanwhile, the earthquakes category has lower performance, suggesting that it is still challenging to classify earthquake-related datasets, and this is likely due to the class imbalance, as seen in Table 1. The model with the highest performance, BERT (16), shows the highest performance in the flood category of 0.793. Followed by earthquake and wildfire scores of 0.756 and 0.750, respectively, the lowest performance was demonstrated in the hurricanes category.

Table 8 shows the performance of various machine learning and deep learning methods using the natural disasters dataset, which consists of four natural disaster datasets. The macro F1 score overall is used to ensure fairness in comparison.

In the transformer model, when compared to the BERT model used in the previous study [11]. The results of this study show that the BERT (16) model, as seen in Figure 7, shows improved performance due to the difference in model architecture, including the addition of specific layers from a previous study [11]. This suggests that differences in model architecture and fine-tuning strategies can impact the final performance. This notion is also reflected in the use of the Random Forest from [7] and [9], Naive Bayes from [9] and [8], and Neural Network from [9] and [10], rely heavily on handcrafted or surface-level feature extraction, which limits their ability to capture complex contextual semantics in disaster-related messages. By leveraging contextualized transformer representations and systematically analyzing fine-tuning strategies, this study advances prior approaches by providing more robust and generalizable performance for eyewitness identification, thereby contributing to state-of-the-art disaster informatics.

Traditional machine learning models are generally limited in understanding semantic context and word order, as they rely on frequency-based and manually engineered feature extraction. Deep learning

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

models have shown improved performance through word embeddings; however, they still face limitations when dealing with social media data, such as high data volume, unstructured text, and noisy words. Therefore, BERT variants based on the Transformer architecture utilize a bidirectional mechanism to capture sentence context from both directions, allowing them to grasp semantic meaning better and handle unstructured social media text. BERT variants also leverage parallel processing compared to previous sequential models, making them more effective in handling the high volume of social media. Additionally, using subword-based tokenization, these models can process noisy words frequently found on social media without losing important information. These approaches address the challenges of social media data in classifying eyewitness disaster-related messages and demonstrate significant performance improvements compared to models used in previous studies.

Table 8. Comparison with Related Works

Source	Method	Macro F1 score (overall)
[7]	Random Forest	0.666
[9]	Random Forest	0.491
[9]	Naïve Bayes	0.503
[8]	Multinomial Naïve Bayes	0.258
[9]	Neural Network	0.485
[10]	Neural Network	0.565
[11]	1D CNN	0.631
[11]	2D CNN	0.630
[11]	3D CNN 1	0.616
[11]	3D CNN 2	0.662
[11]	LSTM	0.602
[11]	BERT	0.578
Our	**BERT (16)**	**0.762**

From an informatics and computer science perspective, this study demonstrates how transformer-based models can achieve robust generalization across multiple disaster categories while remaining scalable for large-scale text processing. The optimized fine-tuning strategy enables efficient deployment without extensive feature engineering, making the approach suitable for high-throughput, real-time text mining scenarios. These properties support the development of automated disaster information systems that can operate reliably under dynamic and time-critical conditions. In practice, this model can be adopted in disaster monitoring or early-warning systems to automatically filter relevant eyewitness reports in real time, reduce manual verification workloads, and support faster situational awareness for emergency responders. It can also support humanitarian dashboards by providing timely, structured information that enhances situational awareness and decision-making for emergency responders.

5. CONCLUSION

This study advances the field of disaster informatics by demonstrating that fine-tuned transformer-based architectures significantly outperform traditional machine learning models for real-time disaster-related text classification. The experimental results show that a batch size of 16 is the optimal configuration, yielding the highest average performance across all natural disaster categories and BERT model variants. In particular, the standard BERT (16) achieved the highest overall macro-F1 score of 0.762, consistently outperforming other variants with and without additional pre-training using Masked Language Modeling

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

(MLM). This highlights the importance of hyperparameter tuning, particularly batch size, in enhancing BERT's effectiveness. By natural disaster category, BERT (16) achieved a macro-F1 score of 0.744 for hurricane, 0.793 for flood, 0.756 for earthquake, and 0.750 for wildfire. Furthermore, BERT (16) significantly improved over machine learning and deep learning models used in previous studies. This suggests that the Transformer architecture, especially with its bidirectional processing, parallel computation, and subword tokenization, offers superior capabilities in capturing semantic context and handling the challenges of unstructured social media text.

Overall, the approach employed in this study successfully improved the classification of eyewitness messages related to natural disasters and supports a more effective response for monitoring of eyewitness messages on social media. For future work, it is recommended to conduct experiments using BERT variants specifically designed for social media text, such as BERTweet, or disaster-domain-oriented models, such as CrisisBERT. These models have been pre-trained on corpora that are more representative of informal language, abbreviations, and noisy writing that are commonly found on social media platforms and disaster-related communications, which may further improve classification performance. Future studies may also explore multilingual or cross-lingual adaptations to support disaster monitoring in diverse linguistic contexts and integrate the proposed approach into real-world disaster monitoring and early warning systems to assess its practical effectiveness. This study highlights the potential of transformer-based models to extract reliable eyewitness information from noisy social media data, supporting timely situational awareness during natural disasters. The findings provide a foundation for developing scalable AI-driven systems that leverage social media for disaster response and other social good applications.

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

[1] V. M. Cvetković, R. Renner, B. Aleksova, and T. Lukić, "Geospatial and Temporal Patterns of Natural and Man-Made (Technological) Disasters (1900–2024): Insights from Different Socio-Economic and Demographic Perspectives," *Applied Sciences*, vol. 14, no. 18, 2024, doi: 10.3390/app14188129.

[2] Y. M. Balakrishna and V. Shivashetty, "Device-to-device based path selection for post disaster communication using hybrid intelligence," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 796–810, Feb. 2024, doi: 10.11591/ijece.v14i1.pp796-810.

[3] R. Efendi and I. R. Widiasari, "Precipitation and water discharge for internet of things based flood disaster prediction improvement," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 6, pp. 6773–6785, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6773-6785.

[4] D. Priyanto, M. Zarlis, H. Mawengkang, and S. Efendi, "Analysis of earthquake hazards prediction with multivariate adaptive regression splines," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2885–2893, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2885-2893.

[5] W. J. Ripple *et al.*, "The 2024 state of the climate report: Perilous times on planet Earth," *Bioscience*, vol. 74, no. 12, pp. 812–824, Dec. 2024, doi: 10.1093/biosci/biae087.

[6] G. Airlangga, "Comparative Analysis of Machine Learning Models for Real-Time Disaster Tweet Classification: Enhancing Emergency Response with Social Media Analytics," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 1, pp. 25–31, 2024.

[7] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Inf Process Manag*, vol. 57, no. 1, p. 102107, Jan. 2020, doi: 10.1016/J.IPM.2019.102107.

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

[8] N. Indrani *et al.*, "Classification of Natural Disaster Reports from Social Media using K-Means SMOTE and Multinomial Naïve Bayes," *J-COSINE (Journal of Computer Science and Informatics Engineering)*, vol. 7, no. ., pp. 60–67, Jun. 2023.

[9] S. Nazir, M. Asif, S. Ahmad, H. Aljuaid, Y. Ghadi, and Z. Nawaz, "Automatic Eyewitness Identification During Disasters by Forming a Feature-Word Dictionary," *Computers, Materials & Continua*, vol. 72, pp. 4755–4769, Nov. 2022, doi: 10.32604/cmc.2022.026145.

[10] S. Haider, M. Azhar, S. Khatoon, M. Alshamari, and M. Afzal, "Automatic Classification of Eyewitness Messages for Disaster Events Using Linguistic Rules and ML/AI Approaches," *Applied Sciences*, vol. 12, pp. 1–17, Oct. 2022, doi: 10.3390/app12199953.

[11] I. Budiman, M. R. Faisal, F. Abadi, D. Nugrahadi, and M. Haekal, "A comparison of word embedding-based extraction feature techniques and deep learning models of natural disaster messages classification," *Journal of Computer Sciences Institute*, pp. 145–153, Jun. 2023, doi: 10.35784/jcsi.3322.

[12] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review," *Algorithms*, vol. 16, no. 5, 2023, doi: 10.3390/a16050236.

[13] Muhammad Zulqarnain *et al.*, "Text Classification Using Deep Learning Models: A Comparative Review," *Cloud Computing and Data Science*, pp. 80–96, Oct. 2023, doi: 10.37256/ccds.5120243528.

[14] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, p. 100157, 2022, doi: https://doi.org/10.1016/j.array.2022.100157.

[15] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," *Information*, vol. 14, no. 4, 2023, doi: 10.3390/info14040242.

[16] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[17] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT," *IEEE Access*, vol. 9, pp. 106363–106374, 2021, doi: 10.1109/ACCESS.2021.3100435.

[18] P. Ganesh *et al.*, "Compressing Large-Scale Transformer-Based Models: A Case Study on BERT," *Trans Assoc Comput Linguist*, vol. 9, pp. 1061–1080, Sep. 2021, doi: 10.1162/tacl_a_00413.

[19] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should you mask 15% in masked language modeling?," *arXiv preprint arXiv:2202.08005*, 2022.

[20] M. Zhao, T. Lin, F. Mi, M. Jaggi, and H. Schütze, "Masking as an efficient alternative to finetuning for pretrained language models," *arXiv preprint arXiv:2004.12406*, 2020.

[21] M. Weyssow, H. Sahraoui, and E. Syriani, "Recommending metamodel concepts during modeling activities with pre-trained language models," *Softw Syst Model*, vol. 21, Dec. 2022, doi: 10.1007/s10270-022-00975-5.

[22] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," *Inf Process Manag*, vol. 59, no. 1, p. 102756, 2022, doi: https://doi.org/10.1016/j.ipm.2021.102756.

[23] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing ELMo and DistilBERT on Socio-political News Classification," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, A. Hürriyetoğlu, E. Yörük, V. Zavarella, and H. Tanev, Eds., Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 9–18. [Online]. Available: https://aclanthology.org/2020.aespen-1.4

[24] K. Clark, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[25] Petr Zelina, "Pretraining and Evaluation of Czech ALBERT Language Model," Masaryk University, Brno, 2020.

[26] V. R. Joseph, "Optimal Ratio for Data Splitting," Feb. 2022, doi: 10.1002/sam.11583.

Vol. 7, No. 3, June 2026, Page. 2307-2322
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5317

[27] M. Naseer, M. Asvial, and R. F. Sari, "An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification," in *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2021, pp. 241–246. doi: 10.1109/ICAIIC51459.2021.9415192.

[28] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," May 2019, [Online]. Available: http://arxiv.org/abs/1905.05583

[29] D. Krstinic, M. Braović, L. Šerić, and D. Božić-Štulić, *Multi-label Classifier Performance Evaluation with Confusion Matrix*. 2020. doi: 10.5121/csit.2020.100801.

[30] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, 2022, doi: 10.1007/s10489-021-02635-5.