Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 1 (2026) 142 - 149 e-ISSN: 2580-0760

# HyRoBERTa: Hybrid Robustly Optimized BERT Approach Model for Sentiment and Sarcasm Detection in Post-Flood Social Media Analysis

Siti Yuliyanti[1], Aveny Septi Astriani[2], Vega Purwayoga[3], Zakwan Gusnadi[4]
[1,3]Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, Indonesia
[2]Department of Indonesian Language and Literature Education, Siliwangi University, Tasikmalaya, Indonesia
[4]Department of Civil Engineering, Faculty of Engineering, Siliwangi University, Tasikmalaya, Indonesia

[1]sitiyuliyanti@unsil.ac.id, [2]aveny.septi@unsil.ac.id, [3]vega.purwayoga@unsil.ac.id, [4]zakwangusnadi@unsil.ac.id

**Abstract**

*The detection problem is a crucial step in sentiment classification because it strengthens the validity and reliability of the model's interpretation of ambiguous text, especially in complex social contexts such as post-disaster public communication. Without this detection, the model is prone to significant classification errors. This study presents a hybrid approach for sentiment analysis with sarcasm detection after a flood disaster by combining the RoBERTa model with sequential deep learning architectures such as GRU, LSTM, and BiLSTM. We used a dataset of 17,520 tweets that were pre-processed using cleaning, normalization, and tokenization. Then, the positive class is further detected to determine whether it is sarcasm. The model was trained using a transformer-based transfer learning method with a combination of hyperparameters: the number of epochs, batch size, dropout rate, and learning rate. The experimental results show that the RoBERTa-GRU model achieved the highest accuracy for sentiment classification at 97. 26%, whereas the RoBERTa-BiLSTM model excels in detecting sarcasm with an accuracy of 98. 74%. RoBERTa-BiLSTM excels in sarcasm detection because it provides a bidirectional sequential mechanism and better long-term memory, effectively leveraging RoBERTa's rich embedding to identify contextual contradictions that are characteristic of sarcasm. Meanwhile, RoBERTa-GRU succeeds in sentiment classification because its architecture is more concise yet effective enough to infer dominant sentiments that have been filtered from the robust representation provided by RoBERTa, making the model more efficient for less complex tasks.*

**Keywords**: hybrid representation model; post floods; RoBERTa, sarcasm; transformers-based

*How to Cite:* S. Yuliyanti, A. Septi Asriani, V. Purwayoga, and Z. Gusnadi, "HyRoBERTa: Hybrid Robustly Optimized BERT Approach Model for Sentiment and Sarcasm Detection in Post-Flood Social Media Analysis ", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 1, pp. 142 - 149, Feb. 2026.
*Permalink*/DOI: https://doi.org/10.29207/resti.v10i1.6963

*Received*: July 15, 2025
*Accepted*: February 11, 2026
*Available Online*: February 25, 2026

*This is an open-access article under the CC BY 4.0 License*
*Published by Ikatan Ahli Informatika Indonesia*

## 1. Introduction

Post-flood disaster problems are the focus of the government due to the mismatch between the needs of the affected victims and the assistance provided. In addition, the delay in evacuating victims is also a problem, but the community is faced with the rapid development of social media as a dissemination of information to express feelings, so that modeling is needed that can be used to analyze post-disaster sentiment towards government or volunteer assistance and logistics so that it is targeted according to the needs of flood victims. Flood disasters will increase in 2025. Floods have a significant impact on human life and the environment. In some areas, flood disasters still cannot be predicted accurately [1], [2]. In disaster management efforts, a quick and accurate understanding of conditions in the field is crucial to maximize response and distribution of aid. Optimization of disaster response can be done by analyzing community responses to flood disasters. Analysis of community responses can be done by utilizing social media [3].

The challenge of Sentiment Analysis in the Flood Context by adding the Role of Sarcasm Detection is crucial because there is a significant barrier, especially in emotional and reactive disaster-related texts: sarcasm. Sarcasm, where the intended meaning contradicts the literal words spoken (e.g., "Amazing, it's already the third day and no help has arrived!"), can

completely skew the results of traditional sentiment analysis. This raises the first issue: the risk of misinterpretation. A sarcastic comment using positive words ("Great," "Good," "Thank you very much") will be classified as positive by the basic sentiment model, when the actual sentiment is negative (frustration, anger).

The second issue is response accuracy. If governments rely on inaccurate sentiment data, they may misjudge the severity of the crisis or the effectiveness of relief efforts. Sarcastic comments are strong indicators of public dissatisfaction and anger, which must be accurately identified to prioritize areas requiring immediate attention.

Social media data has complex characteristics: unstructured, multimodal (text, images, videos), and often ambiguous or subjective[1], [2]. Therefore, a computational approach is needed that is able to manage and analyze such data effectively. Transformer in Natural Language Processing is an attention-based neural network architecture with the concept of encoder-decoder + self-attentio[3], [4]. In this context, transformer models such as BERT, ViT, and CLIP offer great potential through their capabilities in transfer learning, deep semantic representation, and hybrid representation model information integration.

Transformer-based transfer learning enables the utilization of knowledge from general domains to specific domains such as flood disasters, without the need for training from scratch[5]. Meanwhile, the multi-visual representation approach provides a framework for combining information from various types of visual and textual data simultaneously, so that post-flood analysis can be carried out with a wider and more accurate scope[6], [7]. The combination of these two approaches provides an opportunity to build a more adaptive, contextual, and large-scale social media-based flood impact monitoring and evaluation system[8], [9].

Initial research on sentiment analysis and sarcasm detection has been conducted, but there are still many potential research directions to develop. Preliminary studies in sentiment analysis and sarcasm identification have taken place, yet numerous future research avenues remain to be explored. RoBERTa at baseline achieved 84.81% accuracy without sarcasm detection, and 85.09% when sarcasm detection was included—representing only a slight improvement due to the lack of hyperparameter tuning and the hybrid model[10]. RoBERTa + LSTM showed improvement over the baseline, reaching 85.59% without sarcasm detection and 91.37% with sarcasm detection[11]. Previous research shows that IndoBERT and BilLSTM are effective in classifying public sentiment towards natural disasters and climate change issues in Indonesia. The evaluation of the model produced in the study achieved an accuracy of 93.45% and 95.6% with tuning parameters for a learning rate of 0.00002 and a batch size of 16[3]. This indicates that LSTM provides significant added value when sarcasm cues are considered but have not been explored in hyperparameter tuning[12-14]. This suggests that LSTM contributes considerable benefits when handling sarcasm indicators, yet this aspect has not been investigated through hyperparameter adjustment [15]. The ensemble model combining RoBERTa with BiLSTM and GRU achieved high accuracy scores of 91.21% and 91.51%, respectively, with sarcasm detection but did not show significant improvement over the base model[16]. While previous research on sentiment analysis of social media data has focused on general sentiment, few studies have integrated sarcasm detection effectively, a gap this study aims to fill by combining RoBERTa with sequential learning models.

The independent LSTM model exhibited strong performance, achieving a 93.60% accuracy rate, but was exclusively centered on sentiment analysis without incorporating sarcasm detection[10]. The sarcasm detection model, RoBERTa, achieved an impressive accuracy of 98.50%, while DistilBERT followed with 96.20%[17]. These figures highlight the effectiveness of utilizing a transformer-based architecture for sarcasm-sensitive sentiment analysis, but have not yet utilized hybrid models and hyperparameter combinations. These results emphasize the efficiency of employing a transformer-based framework for sentiment analysis that is sensitive to sarcasm, though hybrid models and combinations of hyperparameters have yet to be applied. Based on the gaps in previous research, this study focuses on developing hybrid models: RoBERTA-LSTM, RoBERTa-BiLSTM, and RoBERTa-GRU, with the addition of hyperparameter combinations as a comparative study to improve model accuracy and representation of transformer-based hybrid models using social media data about post-flood disasters. Addressing the shortcomings identified in earlier studies, this research aims to create hybrid models like RoBERTA-LSTM, RoBERTa-BiLSTM, and RoBERTa-GRU, incorporating hyperparameter combinations for a comparative analysis to enhance model accuracy and effectively represent transformer-based hybrid models that utilize social media data regarding post-flood calamities.

This research introduces an innovation unlike earlier studies, specifically by creating a model that merges transformer-assisted transfer learning with combined representation approaches. This model analyzes sentiments expressed on social media following floods, including the detection of sarcasm within positive remarks to ensure that the intended meaning aligns with the interpretation of the public's views on the flood aftermath. The hybrid model incorporates RoBERTa alongside BiLSTM, LSTM, and GRU for sentiment evaluation, followed by sarcasm identification using BiLSTM, which outperforms the other two models. RoBERTa generates embeddings rich in context, while BiLSTM processes the sequence of embeddings by taking into account the broader context from both directions, leading to improved accuracy in identifying

sarcasm. In lengthy, sarcastic remarks, BiLSTM effectively preserves and integrates the meaning from the start to the finish of the sentence.

## 2. Methods

This study presents a hybrid representation of the transfer learning model, utilizing the BERT tokenizer for sentiment classification and the RoBERTa tokenizer for sarcasm detection. The model has undergone text pre-processing, involving a slang dictionary to identify non-standard words or languages. Then split the data, which is continued with classification modeling with the BERT, RoBERTa, and XLM-RoBERTa baselines

RoBERTa acts as a highly sophisticated feature extractor, converting text into a sequence of high-quality vectors. The RNN (GRU/BiLSTM/LSTM) layer then acts as a sequence processor and classifier, extracting fine-grained sequential patterns from the embedding, which subsequently improves the overall accuracy of the model, especially on tasks requiring complex sequence and dependency understanding. RoBERTa excels at understanding context through self-attention (parallel processing), but RNN models excel at modeling sequences (sequential) and long-term dependencies (sequential processing)[3].

While RoBERTa is powerful, the primary reason for combining RoBERTa with RNNs such as GRU, BiLSTM, or LSTM is the hybrid approach.

While RoBERTa-LSTM processes these sequences sequentially, preserving important information across sentences, RoBERTa-GRU's simplicity compared to LSTM makes it an efficient model for less complex tasks (such as straightforward sentiment classification), while RoBERTa-BiLSTM's deep context understanding provides the most significant improvement[16]. BiLSTM processes RoBERTa embedding in two directions (forward and backward) so that after hybrid RoBERTa in sentiment classification, it is more suitable and optimal to use BilLSTM alone for sarcasm detection.

Figure 1 illustrates a Hybrid Representation Model pipeline for post-flood disaster analysis using transformer-based models on social media data.

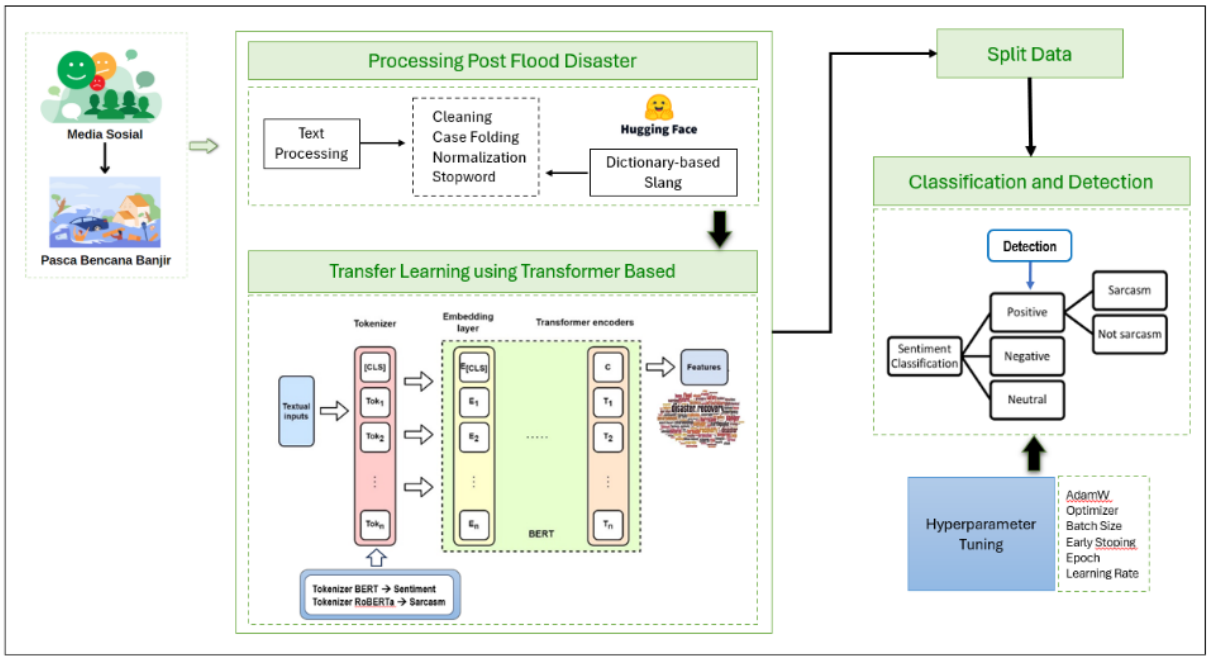



Figure 1. Hybrid Representation Model in Framework Research

Dataset collection stage collects text data about post-flood in 2020 - 2024 from social media X for sentiment analysis of 17,520 tweets shown in Table 1, a slang dictionary for non-standard words or Alay language, stop words using a modified literary dictionary, and an Alay language dictionary or non-standard words for detecting sarcasm class or not when the class enters the positive class.

Text Preprocess stage involves text preprocessing using standard NLP techniques: cleaning (removing noise or irrelevant characters), case folding (converting text to lowercase), stopword removal, and normalization using Slang normalization is done using a dictionary-based approach, leveraging Hugging Face tools.

Text Standardization Informal terms are substituted with conventional equivalents from a predefined dictionary to minimize language diversity[11]. This method is intended to help the model grasp the actual meaning of the text without being swayed by colloquial expressions.

Unwanted symbols are eliminated, yet punctuation such as (!) and (?) as well as (…) are preserved because of their emotional significance[18]. Exclamation points often convey strong feelings like happiness or anger, question marks indicate uncertainty related to emotions like fear or sadness, and ellipses are utilized to signify dramatic pauses, particularly in contexts of sadness and fear.

Table 1. Dataset on flooding in West Java

| ID | Text |
|---|---|
| 1 | Pemkot resmikan Kolam Retensi Gedebage atasi banjir di timur Bandung - ANTARA News Jawa Barat https://t.co/p0zV0wtBF5 |
| 2 | @Assa_yangg Huss gaboleh gt bandung taun lalu banjir lohh gara gara pas taun baru ujannn lebat |
| 3 | @infobdg @MeinNameistMia Rindu lamun Bandung teu heurin ku tangtung teu macet teu banjir aman nyaman teu Aya tindak kriminalitas di jalan... |
| 4 | @PRFMnews @dlhk_kotabdg Sami2 kang supados Bandung teu banjir cileuncang |
| 5 | hiks tahun lalu sekeluarga tahun baruan di bandung. trs tbtb rumah ada yg kena banjir WKWKW pusink gak lo https://t.co/hFDoWYoT9u |
| 6 | Resmikan Kolam Retensi Gedebage Oded M Danial Yakini Bisa Kurangi Banjir Bandung Timur https://t.co/1I62E7exhC |
| … | … |
| 17520 | Sapan udah masuk kab. Bandung. Tp mau misuh suka banjir ekwk makanya kalau temen aku kangen bdg karena muter2 daerah pvj lembang aja aku suka hayang neke |

To create a consistent text format, lowercase conversion is applied, changing all characters to lowercase[2], [19]. This approach is used to remove discrepancies between uppercase and lowercase letters that do not carry specific meanings in emotional analysis.

In this Transfer Learning using Transformer-Based stage, preprocessed textual data is utilized as input for advanced Transformer-based models, specifically BERT and RoBERTa, leveraging the power of transfer learning for downstream classification tasks. The text is first tokenized using appropriate tokenizers tailored to the task: BERT tokenizer is applied for sentiment classification and RoBERTa tokenizer is used for sarcasm detection.

After tokenization, the input tokens are passed through an embedding layer, which transforms tokens into dense vector representations, and multiple transformer encoder layers, which capture contextual relationships between words within the text. The output of this process is a set of contextual features, which serve as the foundation for accurate and robust classification of sentiment and sarcasm in social media posts related to post-flood disaster events.

The models perform two main classification tasks on social media text related to post-flood disaster responses, utilizing pretrained transformer models such as: (1). Sentiment Classification: This task categorizes user posts into three sentiment classes: Positive, Negative, and Neutral [18], [20]. (2). Sarcasm Detection: Sarcasm detection identifies whether a post is sarcastic or not, a task that requires understanding subtle linguistic cues and context[10], [21].

Strongly Enhanced BERT, also known as RoBERTa, shares a very similar structure with BERT [12], [22]. Nonetheless, the developers introduced a few slight modifications to its design and training method to improve the outcomes associated with BERT's framework. We utilized both RoBERTa, employing the twitter-robertabase model that has been trained on approximately 58 million tweets and fine-tuned for sentiment analysis using the TweetEval benchmark, as well as BERT, using the bert-base model from Huggingface [4], [23].

A fundamental component of RoBERTa is self-attention, which can be mathematically defined as in Equation 1.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^2}{\sqrt{dk}})V \quad (1)$$

Q = Query matrix, K = Key matrix, V = Value matrix; and $dk$ = Dimensionality of keys. This explains how the models assign different importance to words in a sentence[11], [24].

Figure 2 illustrates a hybrid deep learning architecture that combines RoBERTa with various types of Recurrent Neural Networks (RNNs) for text classification tasks. In the initial stage of the model pipeline, the input layer processes tokenized text data. This includes two primary components: input_ids and attention_mask. The input_ids represent the numerical encoding of each token in the input text, allowing the model to understand and process the content [25], [26], [27]. Meanwhile, the attention_mask is used to differentiate between actual tokens and padding tokens, ensuring the model focuses only on meaningful parts of the input. Both of these elements are essential for feeding data into the RoBERTa model, enabling it to perform contextual analysis effectively[25].

RoBERTa Layer, a powerful pretrained transformer model, encodes the input tokens into contextual embeddings, capturing deep semantic meaning from the text. These embeddings are then forwarded to three different recurrent branches in parallel: GRU (Gated Recurrent Unit), Bidirectional LSTM (BilLSTM), and LSTM (Long Short-Term Memory) Shown in the Figure 2.

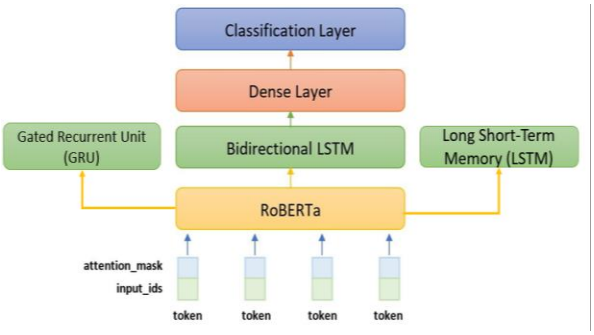



Figure 2. Classification Sentiment Hybrid Representative Architecture

RNN Layers (Sequence Modeling), a unit captures different aspects of temporal dependencies in the sequence: GRU is efficient and fast, useful for shorter or less complex sequences. LSTM handles long-term dependencies better. Bidirectional LSTM captures context from both forward and backward directions of the text, enhancing understanding [13].

Dense Layer (Fusion and Transformation) for outputs from the recurrent layers are likely concatenated or merged and passed through a dense (fully connected) layer to reduce dimensionality and learn deeper abstract representations[28-30]. Classification Layer, the final dense output is passed to the classification layer, which produces the final label prediction sentiment class. After that, if a tweet is positive, it is checked again to detect sarcasm or not with BilLSTM.

These models are fine-tuned on domain-specific data, enabling the system to recognize emotional tones and sarcastic expressions that are common in social media discourse after natural disasters. The model is optimized using several hyperparameters: AdamW Optimizer, Batch Size, Early Stopping, Epochs, and Learning Rate. Evaluation and validation of model performance using a confusion matrix in terms of accuracy, precision, recall, and F1-Score [12], [31].

## 3. Results and Discussions

This research developed various hybrid modeling approaches: RoBERTa-GRU, RoBERTa-BiLSTM, and RoBERTa-LSTM, focused on sentiment analysis tasks. Once the results were classified as positive, they were further validated with BiLSTM to confirm the correctness of the categorization. The initial text processing phase involved cleaning and labeling the data, incorporating polarity to make sure each label's distribution was suitable. The TF-IDF method and tokenization were then utilized for feature extraction. After handling the 17,520 tweets gathered, the dataset was modified to ensure an even distribution of tweets across each category, resulting in 5,840 tweets for positive, negative, and neutral groups through Random Oversampling. Before creating the model, the dataset was split into two parts: a training set and a testing set, with an 80:20 ratio applied to all models. The set parameters consisted of epochs at 10, 50, and 100, batch sizes of 16 and 32, dropout rates of 0.2 and 0.6, along with learning rates of 0.0005 and 0.0001, as well as early stopping. The hyperparameter tuning of the hybrid RoBERTa model using the tested value 64 and the optimizer Adam.

The distribution of model classification in each sentiment class was analyzed using a confusion matrix, which showed that correct predictions in all classes were relatively high, but there were still some prediction errors.

Table 2 shows that the best performance occurred at 5 epochs with a learning rate of 0.0001, achieving 95.54% accuracy. This combination represents the optimal balance between training time and model generalization.

Table 2 showcases the results from the experiments that involved different pairs of learning rates and training epochs, which were assessed based on training loss, validation loss, accuracy in classification, and the time taken for training. The standard RoBERTa model performed its best at 5 epochs with a learning rate set at 0.0001, yielding the highest accuracy rate of 95. 54%, the minimal validation loss of 0.209, and the quickest training duration in the 5-epoch segment of 81.69 seconds. These outcomes suggest that this setup provides an ideal mix of effective learning and generalization capabilities. Additionally, the model that was trained with the lower learning rate of 0.0001 consistently outshone the one with the increased learning rate of 0.0005, especially regarding validation loss and overall accuracy. These results highlight the necessity of precise hyperparameter adjustment to enhance the model's efficiency.

Table 2. The comparative experimental results based on hyperparameters on the baseline RoBERTa model

| Epoch | Learning rates | Train Loss | Val Loss | Accucacy | Time (s) |
|---|---|---|---|---|---|
| 3 | 0.0001 | 0.134 | 0.271 | 94.59% | 82.18 |
| | 0.0005 | 0.102 | 0.219 | 95.34% | 82.01 |
| 5 | 0.0001 | 0.062 | 0.209 | 95.54% | 81.69 |
| | 0.0005 | 0.094 | 0.251 | 95.24% | 81.66 |
| 10 | 0.0001 | 0.137 | 0.255 | 92.67% | 81.15 |
| | 0.0005 | 0.169 | 0.275 | 93.80% | 80.93 |

Table 3 presents a summary of the RoBERTa-BiLSTM model's effectiveness across various combinations of learning rates and training epochs, assessed through metrics such as training loss, validation loss, accuracy, and training duration. The optimal performance was recorded with a learning rate of 0.0001 over 5 training epochs, achieving the highest accuracy of 96. 19% and the lowest validation loss at 0.117, all while keeping the training time efficient at 79. 97 seconds.

Table 3. The comparative experimental results based on the hyperparameters of the RoBERTa-BiLSTM model

| Epoch | Learning rates | Train Loss | Val Loss | Accucacy | Time (s) |
|---|---|---|---|---|---|
| 3 | 0.0001 | 0.133 | 0.242 | 95.84% | 80.15 |
| | 0.0005 | 0.094 | 0.207 | 95.99% | 80.27 |
| 5 | 0.0001 | 0.053 | 0.117 | 96.19% | 79.97 |
| | 0.0005 | 0.093 | 0.202 | 95.89% | 79.99 |
| 10 | 0.0001 | 0.127 | 0.243 | 93.32% | 80.00 |
| | 0.0005 | 0.167 | 0.280 | 94.45% | 80.02 |

This setup reflects a commendable balance between successful learning and the ability to generalize. On the other hand, when the number of epochs was increased to 10, a decline in performance occurred, with accuracy dropping to 93. 32% (for learning rate=0.0001) and 94. 45% (for learning rate=0.0005), along with a rise in validation loss, suggesting possible overfitting. In every scenario, models that utilized the lower learning rate of 0. 0001 demonstrated consistently superior validation results, highlighting the significance of selecting an appropriate learning rate. These findings emphasize that fine-tuning hyperparameters, especially regarding the number of epochs and learning rate, is essential for enhancing the proposed model's performance.

Table 4 displays the evaluation metrics for the RoBERTa-LSTM model across different combinations of epochs and learning rates. The metrics assessed include training loss, validation loss, accuracy, and the time taken for training. The findings reveal that the best

performance occurred at the third epoch with a learning rate of 0. 0001, which resulted in the lowest training loss (0. 014) and validation loss (0. 026), as well as the highest accuracy (96. 54%) and a training duration of 79. 00 seconds. In contrast, other settings led to either greater losses or lower accuracy, even though the training times were similar. It is worth noting that increasing the epochs to 10 did not improve performance and, in some instances, caused a decline in both loss metrics and accuracy. These results indicate that implementing early stopping with a suitable learning rate can effectively enhance model performance while reducing computational expenses.

Table 4. The comparative experimental results based on hyperparameters of the RoBERTa-LSTM model

| Epoch | Learning rates | Train Loss | Val Loss | Accuracy | Time (s) |
|---|---|---|---|---|---|
| 3 | 0.0001 | 0.112 | 0.202 | 92.06% | 79.17 |
| | 0.0005 | 0.073 | 0.140 | 94.01% | 79.30 |
| 5 | 0.0001 | 0.014 | 0.026 | 96.54% | 79.00 |
| | 0.0005 | 0.066 | 0.191 | 96.21% | 79.02 |
| 10 | 0.0001 | 0.119 | 0.201 | 94.07% | 78.53 |
| | 0.0005 | 0.159 | 0.221 | 93.28% | 78.54 |

Table 5 outlines the results of experiments comparing the RoBERTa-GRU model under various settings of epochs and learning rates. The best results were found at epoch 5 with a learning rate of 0. 0001, resulting in the lowest training loss of 0. 013 and validation loss of 0. 018, as well as the highest accuracy of 97. 26%, all achieved in a training period of 78. 58 seconds. Other setups, such as epoch 5 with a learning rate of 0. 0005, also yielded impressive accuracy at 96. 93%. In summary, the results emphasize the importance of adjusting the learning rate and choosing the right number of epochs to enhance the RoBERTa-GRU model's performance while keeping computational demands low.

Table 5. The comparative experimental results based on hyperparameters of the RoBERTa-GRU model

| Epoch | Learning rates | Train Loss | Val Loss | Accucacy | Time (s) |
|---|---|---|---|---|---|
| 3 | 0.0001 | 0.093 | 0.105 | 92.78% | 78.75 |
| | 0.0005 | 0.054 | 0.065 | 94.73% | 78.88 |
| 5 | 0.0001 | 0.013 | 0.018 | 97.26% | 78.58 |
| | 0.0005 | 0.047 | 0.093 | 96.93% | 78.60 |
| 10 | 0.0001 | 0.100 | 0.104 | 94.89% | 78.11 |
| | 0.0005 | 0.140 | 0.124 | 94.15% | 78.12 |

Next, the evaluation of the model's effectiveness in identifying sarcasm among all individuals in the positive class is conducted to confirm that the positive class does not represent sarcasm. This utilizes the top model in classification, specifically with 5 epochs and a learning rate of 0.0001, as indicated in the Table 5.

Table 6 presents a comparative analysis of sentiment classification accuracy using different model configurations, both with and without sarcasm detection. The results are based on the best-performing hyperparameter setting (epoch = 5, learning rate = 0. 0001). The baseline RoBERTa model achieved an accuracy of 95. 54% with sarcasm detection and 96. 81% without sarcasm detection. Incorporating additional sequence modeling layers significantly improved performance. Specifically, RoBERTa + BiLSTM attained the highest accuracy of 98. 74% without sarcasm detection and 96. 19% with sarcasm detection, followed closely by RoBERTa + LSTM, which achieved 97. 96% and 96. 54%, respectively. The RoBERTa + GRU configuration yielded the highest accuracy with sarcasm detection at 97. 26%, and 98. 26% without sarcasm detection, indicating that GRU offers a strong balance between sarcasm-aware and sarcasm-agnostic performance. Overall, the integration of sequential neural layers such as GRU, LSTM, and BiLSTM with RoBERTa enhances the model's ability to handle nuanced sentiment, particularly in the presence of sarcasm.

Table 6. The comparative experimental results based on sarcasm detection with the best model (epoch = 5 and learning rate = 0.0001)

| Model | Sentiment Classification Accuracy | |
|---|---|---|
| | Using sarcasm detection | Without sarcasm detection |
| RoBERTa | 95.54% | 96.81% |
| RoBERTa + BilLSTM | 96.19% | 98.74% |
| RoBERTa + LSTM | 96.54% | 97.96% |
| RoBERTa + GRU | 97.26% | 98.26% |

The model performance after sarcasm detection shows the best hybrid model order, namely from RoBERTa-BiLSTM, RoBERTa-LSTM, and finally RoBERTa-GRU, which is different from the sentiment classification results, which are outperformed by RoBERTa-GRU. RoBERTa-BiLSTM is superior to RoBERTa-GRU because BiLSTM can capture two-way context, which is very important in detecting sarcasm, where the meaning of a sentence is often hidden at the end of a sentence or reversed from the beginning. To ensure that the results of this study are superior to previous research, Table 7 contains a comparison.

Table 7 outlines a comparison of sentiment classification accuracy between earlier models and the newly suggested hybrid models, considering both scenarios with and without sarcasm detection. The findings reveal that prior models, like RoBERTa and RoBERTa plus LSTM, recorded accuracies of 84. 81% and 85. 59%, respectively, in the absence of sarcasm detection. These figures improved to 85. 09% and 91. 37% when sarcasm detection was incorporated. Ensemble methods, such as RoBERTa paired with BiLSTM or GRU, also demonstrated enhanced baseline accuracies of 91. 21% and 91. 51%, although they did not share results pertaining to sarcasm detection. Likewise, the LSTM model reached an accuracy of 93. 60% without managing sarcasm. In contrast, studies concentrating on improving sarcasm detection showed significant progress, with RoBERTa and DistilBERT achieving accuracies of 98. 50% and 96. 20% respectively.

The research proposed here surpasses previous approaches by utilizing a hybrid model alongside systematic hyperparameter tuning. Among the models

put forward, the RoBERTa + BiLSTM combination attained the top accuracy of 96. 19% without sarcasm detection and 98. 74% when sarcasm detection was applied. Other hybrid setups, such as RoBERTa + LSTM and RoBERTa + GRU, also demonstrated strong performance, validating the proposed framework's effectiveness in improving sentiment classification, especially when sarcasm is accurately identified and managed.

Table 7. Comparison of model performance results with previous research

| Model | | Accuracy Sentiment Classification | |
|---|---|---|---|
| | | Without Sarcasm Detection | Using Sarcasm Detection |
| RoBERTa [24] | | 84.81% | 85.09% |
| RoBERTa + LSTM [14] | | 85.59% | 91.37% |
| Ensemble [3] | RoBERTa + BiLSTM | 91.21% | - |
| | RoBERTa + GRU | 91.51% | - |
| LSTM [16] | | 93.60% | - |
| Improve Detection of Sarcasm [11] | RoBERTa | - | 98.50% |
| | DistilBERT | - | 96.20% |
| Our Research (comparative hyperparameter, hybrid model using sarcasm detection) | RoBERTa | 95.54% | 96.81% |
| | RoBERTa + BilLSTM | 96.19% | 98.74% |
| | RoBERTa + LSTM | 96.54% | 97.96% |
| | RoBERTa + GRU | 97.26% | 98.26% |

## 4. Conclusions

This research proposed and evaluated hybrid models for sentiment analysis of post-flood disaster tweets by integrating the RoBERTa transformer with sequential deep learning architectures (GRU, LSTM, and BiLSTM), enhanced through comprehensive hyperparameter tuning and sarcasm detection.

The challenge of sentiment analysis in the context of flood disasters is crucial, especially in the context of sarcasm detection. Sarcasm detection poses significant obstacles to sentiment classification, especially in emotional and reactive disaster-related texts that contain sarcastic content, potentially leading to misinterpretation, low response accuracy, and the identification of contradictions between lexicons (literal words).

Key findings from hybrid model performance is among the three hybrids, RoBERTa-GRU achieved the highest sentiment classification accuracy of 97.26%, outperforming other combinations in sentiment classification tasks. However, RoBERTa-BiLSTM demonstrated the best sarcasm detection performance 98.74%, highlighting its superior ability to interpret bidirectional context—an essential aspect of detecting sarcasm. Effectiveness of Sarcasm detection enhanced classification robustness, particularly for positive sentiments, where sarcastic statements can be misclassified without contextual analysis.

Hyperparameter tuning insights result the optimal setup across all models was consistently found with 5 training epochs and a learning rate of 0.0001, balancing accuracy, generalization, and training efficiency. Comparison with previous research the proposed models surpassed all benchmarked studies, including earlier RoBERTa-only and ensemble-based methods. RoBERTa + BiLSTM achieved the highest overall accuracy 98.74% with sarcasm detection, demonstrating the effectiveness of combining contextual transformers with sequence-aware deep learning.

This hybrid model can be a valuable tool for real-time sentiment analysis in post-disaster scenarios, helping agencies monitor public opinion and adjust their response strategies accordingly. Recommendations for further research include implementing cross-lingual sentiment classification because some tweets in the dataset have a mixture of languages, and then deploying it in the form of an application so that users can adjust hyperparameters according to model needs.

## Acknowledgements

The Research Team would like to express their gratitude to DPPM Ditjen Risbang Kemdiktisaintek for funding this research in the 2025 fiscal year under Grant Contract 107/C3/DT.05.00/PL/2025, and LPPM Siliwangi University for assisting with administrative completeness from preparation until this research can be completed.

## References

[1] S. Yuliyanti and S. Sholihah, “Pemodelan Analisis Sentimen Masyarakat terhadap Adaptasi Kebiasaan Baru (AKB) mengunakan Algoritma Naïve Bayes,” *MIND J.*, vol. 6, no. 2, pp. 155–167, 2021, doi: 10.26760/mindjournal.v6i2.155-167.

[2] S. Yuliyanti, T. Djatna, and H. Sukoco, “Sentiment mining of community development program evaluation based on social media,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 4, pp. 1858–1864, 2017, doi: 10.12928/TELKOMNIKA.v15i4.4633.

[3] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, “Sentiment Analysis With Ensemble Hybrid Deep Learning Model,” *IEEE Access*, vol. 10, pp. 103694–103704, 2022, doi: 10.1109/ACCESS.2022.3210182.

[4] T. Bey Kusuma, I. Komang, and A. Mogi, “Implementasi BERT pada Analisis Sentimen Ulasan Destinasi Wisata Bali,” *J. Elektron. Ilmu Komput. Udayana*, vol. 12, no. 2, pp. 409–420, 2023 https://doi.org/10.51519/journalisi.v6i3.814.

[5] N. Singh, N. Roy, and A. Gangopadhyay, “Localized Flood DetectionWith Minimal Labeled Social Media Data Using Transfer Learning,” 2020, [Online]. Available: https://arxiv.org/abs/2003.04973.

[6] Y. Tasya and R. A. Putri, “Klasifikasi Tingkat Kerawanan Banjir Wilayah Medan Menggunakan Metode Naive Bayes Dan Algoritma J48,” *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 657–665, 2023, doi: 10.31539/intecoms.v6i2.7392.

[7] B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, and A. Mosavi, “An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines,” *Sci. Total Environ.*, vol. 651, pp. 2087–2096, Feb. 2019, doi: 10.1016/j.scitotenv.2018.10.064.

[8] A. M. Maksun, Y. A. Sari, and B. Rahayudi, “Analisis Sentimen pada Twitter Bencana Alam di Kalimantan Selatan

menggunakan Metode Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 12, pp. 5614–5621, 2021, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10307

[9] R. Dhawale, C. J. Schuster-Wallace, and A. Pietroniro, "Assessing the multidimensional nature of flood and drought vulnerability index: A systematic review of literature," *Int. J. Disaster Risk Reduct.*, vol. 112, no. August, p. 104764, 2024, doi: 10.1016/j.ijdrr.2024.104764.

[10] T. Dadu and K. Pant, "Sarcasm detection using context separators in online discourse," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 51–55, 2020, doi: 10.18653/v1/P17.

[11] P. Dubey, P. Dubey, and P. N. Bokoro, "Unpacking Sarcasm: A Contextual and Transformer-Based Approach for Improved Detection," *Computers*, vol. 14, no. 3, pp. 1–21, 2025, doi: 10.3390/computers14030095.

[12] F. Ullah, A. Alsirhani, M. M. Alshahrani, A. Alomari, H. Naeem, and S. A. Shah, "Explainable Malware Detection System Using Transformers-Based Transfer Learning and Multi-Model Visual Representation," *Sensors*, vol. 22, no. 18, 2022, doi: 10.3390/s22186766.

[13] R. H. Yahya, W. Maharani, and R. Wijaya, "Disaster Management Sentiment Analysis Using the BiLSTM Method," *J. Media Inform. Budidarma*, vol. 7, no. 1, p. 501, 2023, doi: 10.30865/mib.v7i1.5573.

[14] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[15] M. Chiny, M. Chihab, and Y. Chihab, "LSTM , VADER and TF-IDF based Hybrid Sentiment Analysis Model," vol. 12, no. 7, pp. 265–275, 2021. Doi: 10.14569/IJACSA.2021.0120730.

[16] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," *Sci. Rep.*, vol. 14, no. 1, p. 24882, 2024, doi: 10.1038/s41598-024-76079-5.

[17] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, 2022, doi: 10.3390/s22218184.

[18] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA and BERT Model.," *Procedia Comput. Sci.*, vol. 218, no. 2022, pp. 908–917, 2022, doi: 10.1016/j.procs.2023.01.071.

[19] N. Yusliani, A. Nabilah, M. Raihan Habibullah, A. Darmawahyuni, and G. Athalina, "Comparing Word Representation BERT and RoBERTa in Keyphrase Extraction using TgGAT," *J. RESTI*, vol. 9, no. 2, pp. 250–257, 2025, doi: 10.29207/resti.v9i2.6279.

[20] G. Hakim, T. N. Fatyanosa, and A. W. Widodo, "Analisis Sentimen Masyarakat terhadap Kereta Cepat Whoosh pada Platform X menggunakan IndoBERT," vol. 1, no. 1, pp. 1–10, 2023, https://doi.org/10.24076/joism.2026v7i2.2449.

[21] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "CASCADE: Contextual sarcasm detection in online discussion forums," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, pp. 1837–1848, 2018 https://aclanthology.org/volumes/C18-1/.

[22] T. Wahyuningsih *et al.*, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 2, pp. 343–348, 2021, doi: 10.17762/turcomat.v12i3.938.

[23] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile, "AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets," *CEUR Workshop Proc.*, vol. 2481, 2019. https://ceur-ws.org/Vol-2481/paper57.pdf.

[24] J. Nasional, S. Informasi, E. Teks, E. Marthen, and I. Novita, "Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi," vol. 03, pp. 240–248, 2024. https://doi.org/10.52436/1.jutif.2025.6.5.4897.

[25] S. Informasi, F. Teknik, and U. Bakrie, "Perbandingan Performa Algoritma VADER dan RoBERTa pada Twitter Comparison of VADER and RoBERTa Algorithm Performance on," vol. 13, pp. 1547–1557, 2024.

[26] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 1, p. 79, 2023, doi: 10.22146/ijccs.77354.

[27] F. Sufi, "A Sustainable Way Forward: Systematic Review of Transformer Technology in Social-Media-Based Disaster Analytics," *Sustain.* , vol. 16, no. 7, 2024, doi: 10.3390/su16072742.

[28] W. Chen *et al.*, "Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles," *J. Hydrol.*, vol. 575, pp. 864–873, Aug. 2019, doi: 10.1016/j.jhydrol.2019.05.089.

[29] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11856 LNAI, no. 2, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[30] A. U. Khan, A. Mazaheri, N. da Vitoria Lobo, and M. Shah, "MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020*, pp. 4648–4660, 2020, doi: 10.18653/v1/2020.findings-emnlp.417.

[31] M. G. Villar, J. B. Ballester, I. De, T. Diez, and I. Ashraf, "Analyzing Sentiments Regarding ChatGPT Using Novel BERT : A Machine Learning Approach," pp. 1–29, 2023. https://www.mdpi.com/2078-2489/14/9/474.