SentiBERT and Enhanced Bi-GRU for Weather-related Text Classification Using Lexical Features

Mohamad Anwar Syaefudin [1,*)], Arijal Ibnu Jati [2)], Hilya Tsaniya [3)], Chastine Fatichah [4)], and Diana Purwitasari [5)]

[1, 2, 3, 4, 5)] Department of Informatics, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

E-mail: 6025232002@student.its.ac.id[1)], 6025241044@student.its.ac.id[2)], 7025222004@student.its.ac.id[3)], chastine@its.ac.id[4)], and diana@its.ac.id[5)]

ABSTRACT

The growing volume of Weather-related content on social media platforms, especially Twitter, has highlighted the need for robust classification models that can handle noisy, ambiguous, and emotionally subtle language. However, existing models machine learning such as Support Vector Machines (SVM) often fail to effectively capture implicit sentiment and sequential context in short, real time texts. This study addresses the challenge of Weather-related text classification by proposing a hybrid architecture that combines SentiBERT, a sentiment aware transformer model, with an Enhanced Bi-GRU network equipped with Self Attention and LeakyReLU activation. Experiments were conducted using a five class(sunny, cloudy, rainy, extreme, other) dataset of Weather-related tweets with stratified cross validation across multiple deep learning models and tokenizers. Results show that the proposed SentiBERT + Enhanced Bi-GRU model outperformed all baselines, achieving 88.03% accuracy and 88.25% macro F1 score demonstrating its ability to better interpret contextual and emotional nuances. These findings imply that integrating sentiment specific embeddings with sequential modeling and lexical features offers a promising direction for future real time applications in climate monitoring and disaster alert systems.

Keywords: Weather text classification, SentiBERT, Bi-GRU, self attention, lexical features, sentiment analysis.

1. Introduction

As an archipelagic country, Indonesia possesses highly diverse geographical and ecological conditions. The combination of a large population, high biodiversity, a coastline stretching over 80,000 km, and more than 17,000 islands makes Indonesia one of the most vulnerable countries to climate change [1]. To understand the impacts of climate change in Indonesia, it is essential to first recognize its climatic characteristics. Indonesia experiences two distinct seasons, the rainy (wet) season and the dry season, with fluctuating levels of rainfall [2]. These frequently changing weather conditions encourage the public to seek and share weather-related information, one of which is through social media platforms like Twitter, which has become an important source for monitoring daily weather events. Leveraging Twitter data allows for operational support during extreme weather events via local forecasting, surveillance, and impact monitoring [3]. The increasing use of social media to share weather information has made text mining techniques increasingly important for analyzing text-based data obtained from microblogs like Twitter. Text mining from microblogs such as Twitter is one method to gather information about weather events from various text sources [4]. Natural Language Processing (NLP) can enhance weather event assessments by analyzing data sources such as social media and news articles, generating new information, and reducing monitoring costs [5]. Text classification constitutes a core and essential operation in the field of natural language processing [6]. Within the realm of weather event analysis utilizing social media data, text classification serves as a pivotal technique for organizing and interpreting the extracted textual information. By employing text classification algorithms, raw text data from platforms like Twitter can be systematically categorized based on relevant criteria such as the type

* Corresponding author.
Received: June 19th, 2025. Revised: October 10th, 2025. Accepted: December 1st, 2025.
Available online: January 15th, 2026.

DOI: https://doi.org/10.12962/j24068535.v24i1.a1320

of weather phenomenon reported, the perceived sentiment towards the event, or the urgency of the message. This categorization facilitates subsequent in-depth analysis and informed decision-making in disaster management and situational awareness. Several machine learning approaches, such as Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression (LR), have been used to classify weather-related texts on Twitter. Results from previous studies show that the SVM algorithm provides the most accurate results, with an accuracy rate of up to 93% in classifying weather conditions based on Twitter data [7], [8]. Pretrained language models like BERT have proven to outperform traditional machine learning methods, achieving high F1-scores in weather-related text classification tasks [9]. FF-BERT, a BERT-based ensemble model combining CNN and bagging, improved micro-F1 by 11.83% over baseline, highlighting the strength of pretrained models with ensemble learning for classifying flash flood-related texts [10]. Similarly, Debata and Elango (2024) proposed a multimodal flood prediction model integrating BERT, LSTM, and FastText within an ensemble framework, which, when combined with meteorological data, achieved an accuracy of 97%, demonstrating the effectiveness of combining statistical data and social media analysis for real-time flood forecasting [11].

1.1. Research Gap

However, while BERT-based models have shown remarkable performance in general sentiment analysis, they often fall short in capturing subtle, implicit, or contextually nuanced sentiments commonly found in weather-related social media texts. This is partly due to BERT's tendency to struggle with detecting sentiments that are not explicitly stated, or those that are embedded beneath the surface meaning of words, which often leads to reduced accuracy when dealing with ambiguous or implied emotional expressions [12], [13]. In the context of weather-related tweets, users frequently express concern, urgency, or uncertainty in indirect ways, such as sarcastic remarks, rhetorical questions, or emotionally neutral terms that carry deeper situational meaning. These forms of sentiment are challenging to identify using standard pretrained language models, as they require deeper contextual understanding and sensitivity to linguistic subtleties. Moreover, the noisy and short nature of social media texts introduces additional complexity in interpreting sentiment signals accurately. As a result, models that lack mechanisms to handle these nuances often underperform in real-world, domain-specific scenarios like weather-related text classification. Although recent studies have explored hybrid models combining Transformers with sequential layers like Bi-GRU and enhancements using lexical features, few works have specifically integrated sentiment-aware models such as SentiBERT with Enhanced Bi-GRU alongside lexical-level enhancements (e.g., polarity and subjectivity), particularly for Indonesian weather-related texts on social media.

1.2. Contribution

To address these challenges, this paper proposes a hybrid framework that combines SentiBERT, a sentiment-aware language model, with an Enhanced Bi-GRU architecture equipped with Self-Attention and LeakyReLU activation, along with lexical features such as polarity and subjectivity, for the classification of weather-related social media texts. The use of SentiBERT is motivated by its ability to better capture sentiment-oriented representations compared to general-purpose models like BERT, especially in contexts where emotional cues are subtle or implicit. Meanwhile, Bi-GRU is chosen for its effectiveness in modeling sequential dependencies in short and noisy texts. The incorporation of lexical features provides complementary sentiment signals that reinforce the model's understanding. The contributions of this paper are summarized as follows: (1) Proposes a novel hybrid model integrating SentiBERT and Enhanced Bi-GRU with sentiment-oriented lexical features, specifically tailored for multi-class weather-related text classification (sunny, cloudy, rainy, extreme, other) in Indonesian social media data. (2) Conducts a comprehensive empirical comparison across multiple transformer-based models (e.g., BERT, RoBERTa, DistilBERT) and deep learning architectures (RNN, CNN, Transformer), demonstrating superior performance in accuracy (88.03%) and macro-F1 score (88.25%) for real-time climate monitoring and disaster alert applications.

2. Related Work

Weather-related text classification has become a crucial area of research due to the increasing volume of real time data available on social media platforms. Early studies utilized traditional machine learning techniques like

Support Vector Machines (SVM), Multinomial Naive Bayes, and Logistic Regression to classify Weather-related content. For instance, Purwandari et al. (2023) reported that SVM achieved up to 93% accuracy in classifying weather-related tweets using explicit keyword-based features [7], [8]. However, traditional models like SVM often struggle to capture implicit sentiment, ambiguity, or sequential dependencies in short, noisy texts, limitations well-documented in prior studies [12], [13].

With the rise of more complex language models, the use of BERT (Bidirectional Encoder Representations from Transformers) has become prominent in the field of Weather-related text classification. Anoop et al. (2023) fine tuned ClimateBERT, a BERT variant trained on climate related texts, and showed that it outperformed traditional models in capturing nuanced sentiments about climate change and weather events. This approach emphasized the power of domain specific pre trained models to understand the unique language of Weather-related discourse [14]. Effrosynidis et al. (2022) demonstrated that sentiment analysis on climate change related tweets, using techniques like BERT, significantly improved the classification of stance and sentiment compared to traditional models [15]. Their findings reveal that BERT's ability to process contextual information allows for better classification of user opinions on climate issues, further reinforcing the importance of domain-specific models for environmental discourse.

Building on these advancements, hybrid models combining BERT with sequential models like Bi-GRU have emerged as a promising direction for improving Weather-related text classification. Tan et al. (2023) proposed a RoBERTa-GRU model, leveraging the contextual embeddings of RoBERTa with the sequential power of GRU, showing significant improvements in sentiment analysis, particularly when dealing with imbalanced datasets [16]. Their research highlighted how combining Transformer based models with sequence models can better capture both the contextual and sequential dependencies inherent in Weather-related texts. Similarly, Effrosynidis et al. (2022) demonstrated that BERT achieved the highest accuracy in classifying sentiment, stance, and aggressiveness in climate-related tweets, outperforming traditional models like SVM, CNN, and LSTM [17]. Their results highlight the superior performance of transformer-based models in extracting nuanced information from weather and climate-related texts.

Incorporating lexical features, such as word intensity and negation, has also been explored to enhance classification accuracy. Lei Shi et al. (2022) demonstrated that integrating lexical features with BERT could improve disaster related text classification, showing that these features provide valuable insights that contextual embeddings alone may miss [18]. Similarly, models like SentiBERT, which fine tunes BERT to capture sentiment specific embeddings, have proven effective in analyzing public sentiment during extreme weather events.

Research related to Weather-related text classification has shown significant advancement with the use of transformer models such as BERT, RoBERTa, and ClimateBERT. These models effectively capture contextual meaning but often struggle with sequential dependencies. To address this, recent studies have explored hybrid models combining Transformers with sequential layers like Bi-GRU, and enhancements using lexical features. Sentiment-aware models such as SentiBERT have further improved sentiment classification, particularly during extreme weather events. However, few works have explored combining SentiBERT with Bi-GRU alongside lexical level enhancements. Our study addresses this gap by integrating SentiBERT with an enhanced Bi-GRU architecture using Self Attention and LeakyReLU, resulting in significant performance gains across all evaluation metrics.

3. Methodology

The methodology for this study is carefully constructed to facilitate the development and evaluation of a weather-related text classification model using historical Twitter data. To achieve this, the proposed strategy innovatively integrates SentiBERT with an Enhanced Bi-GRU classifier. This novel approach is subsequently compared against various other language models, including BERT, RoBERTa, and DistilBERT, each paired with distinct neural network architectures such as RNN, CNN, and Transformer. The workflow comprises several critical stages: Data Collection, Data Preprocessing, Tokenization and Embedding, Classification, and Performance Evaluation, as illustrated in Fig. 1. The implementation prioritizes comprehensive analysis and comparison to ensure robustness

and accuracy in classifying intricate weather-related textual data, reflecting the dynamics of real-world conditions often expressed ambiguously in social media communications.

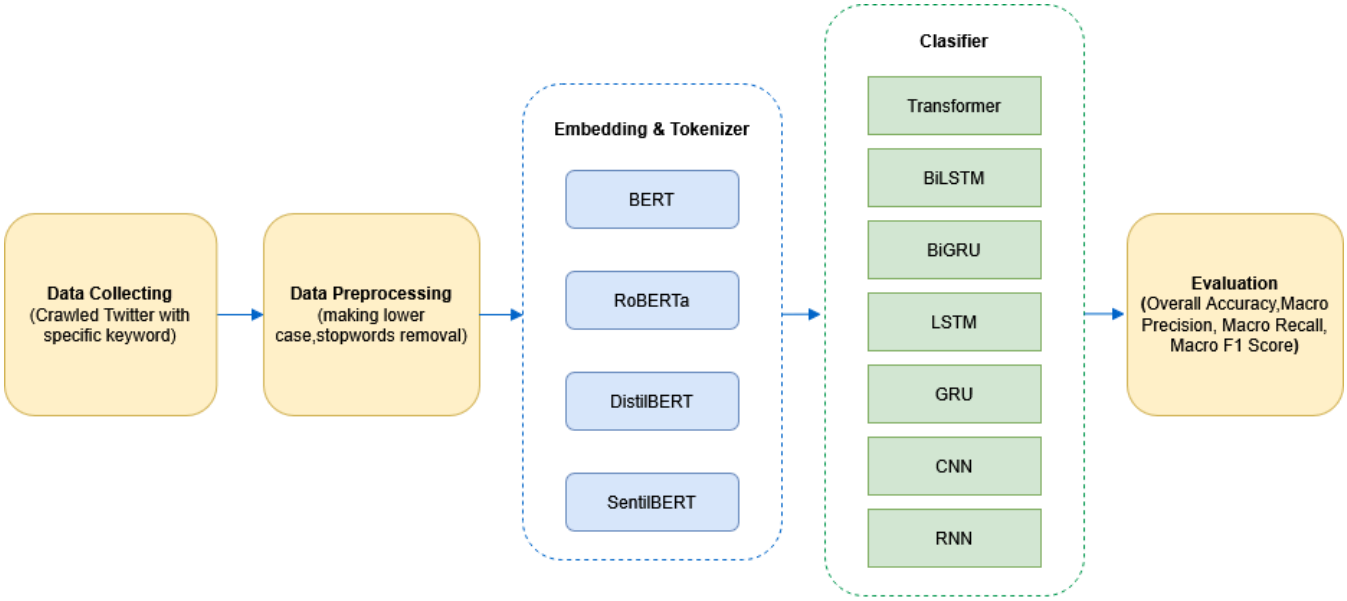

Fig. 1: Flowchart for weather-related text classification.

All sentences are first converted into token representations (word embeddings), which are then processed through deep learning-based classifiers. Three different models are used: RNN-based, CNN-based, and Transformer-based. In the RNN-based model, embeddings are passed through sequential layers such as GRU or Bi-GRU, followed by a multi-layer perceptron (MLP) with linear layers and ReLU activation to map features into class labels. The CNN-based model applies multiple convolution filters with varying kernel sizes to extract local patterns, followed by ReLU activation, max pooling, and dense layers for classification. Meanwhile, the Transformer-based model utilizes self-attention by computing query, key, and value vectors to generate attention outputs, which are aggregated via mean pooling before being passed to the final classification layer. Each model captures different aspects of linguistic patterns, allowing a comprehensive evaluation of their effectiveness in weather-related tweet classification.

3.1. Dataset

Before applying the training model to the dataset, we conducted a preliminary analysis to better understand the data characteristics and potential challenges. The dataset used is non public data obtained by crawling Twitter using a set of weather-related keywords, resulting in a total of 797 Indonesian language sentences that reflect real world expressions of weather phenomena. The dataset is imbalanced with the following class distributions: sunny (106), cloudy (131), rainy (172), extreme (95), and other (293). The other class has the largest number of samples, highlighting key challenge in this experiment, how to appropriately incorporate this class into the model. We use the manual labeling process, only a single annotator was involved. The correction process was conducted through a trial-and-error approach by running the dataset using one model for a single epoch. If the resulting performance metrics (e.g., Accuracy, F1-score) were not satisfactory, the labeling was revised and the model was rerun using the corrected dataset. This procedure was repeated until the dataset that produced the most optimal results was selected.

The ‘other’ class plays a crucial role in this experiment because, in real world weather data, many sentences do not clearly align with the four primary categories (sunny, cloudy, rainy, extreme). Previous studies often excluded such data by focusing solely on these main classes. We hypothesize that incorporating the ‘other’ class can enhance the model’s performance, particularly in handling ambiguous or difficult to classify sentences. Grouping non specific data into this class enables the model to better learn the characteristics of each main class, thereby benefiting classes like ‘extreme’ that have fewer data samples. This approach may also contribute to reducing overfitting.

Additionally, a word cloud was generated for each class to support our analysis. A word cloud provides a visual representation of the most frequent or significant words in the dataset, with commonly occurring words shown in larger font sizes or more prominent colors. These word clouds, presented in Fig. 2, help illustrate the linguistic patterns associated with each class. These visualizations not only reveal dominant lexical items but also assist in identifying overlapping vocabulary across categories such as the frequent appearance of the word “hujan” (rain) in multiple classes, which emphasizes the importance of context-aware models.

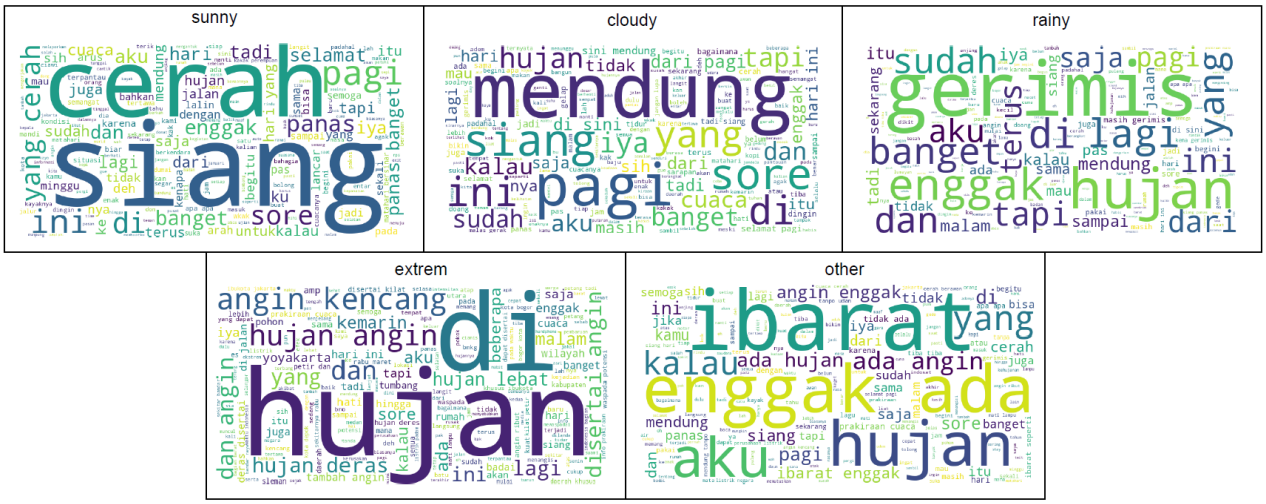

Fig. 2: Word cloud for each class.

From the word cloud visualization shown in Fig. 2, it can be observed that each class used in the experiment contains distinct lexical patterns, where the word “hujan” (rain) appears not only in the rainy class but also in other classes such as cloudy, extreme, and even other. This phenomenon indicates that weather-related words like “hujan” have broad semantic associations and can appear in various contextual meanings. For example, in a tweet such as “hujan deras disertai angin kencang” (heavy rain accompanied by strong winds), although the word “hujan” is present, the overall context is more relevant to the extreme category. On the other hand, a sentence like “mendung banget, kayaknya mau hujan” (very cloudy, looks like it’s going to rain) would likely fall under the cloudy class, despite the mention of “hujan”. The presence of “hujan” in the other class further shows that the existence of a weather keyword alone does not necessarily define the complete meaning of the text, especially when no clear or dominant weather condition is described. This emphasizes the importance of models that go beyond simple keyword detection and are capable of capturing contextual and emotional nuances within the text. In this regard, the SentiBERT model combined with Enhanced Bi-GRU demonstrates an advantage in understanding both contextual representations and implicit sentiment. Therefore, the word cloud visualization supports the relevance of using a hybrid semantic-sequential approach in weather-related text classification on social media, which is often ambiguous and not explicitly stated.

3.2. SentiBERT

3.2.1. Overview of SentiBERT

Sentiment analysis remains a central task in natural language processing, with compositional semantics posing a persistent challenge, particularly in the presence of negation and contrast (e.g., “Frenetic but not really funny”). Early efforts employed recursive neural networks over constituency trees to model how token and phrase meanings combine (Socher et al., 2012, 2013), however such approaches often paid insufficient attention to wider contextual associations. This paper [cite] introduce SentiBERT, which integrates BERT-based contextual embeddings (Devlin et al., 2019; Liu et al., 2019) with a recursive tree structure, enhanced by a two-stage attention network that composes sentiment representations at each phrase node. Training combines masked language modeling with phrase-level sentiment supervision on the Stanford Sentiment Treebank, yielding significant improvements over both recursive

baselines and vanilla BERT in phrase-level classification. Moreover, the learned compositional semantics transfer effectively to related tasks Twitter sentiment analysis, emotion-intensity prediction, and contextual emotion detection. Comprehensive quantitative and qualitative analyses further demonstrate SentiBERT ability to capture nuanced sentiment composition. The implementation is publicly accessible at https://github.com/WadeYin9712/SentiBERT.

These characteristics make SentiBERT particularly suitable for analyzing short, noisy, and emotionally ambiguous texts such as tweets. Its ability to model phrase-level sentiment allows it to go beyond surface-level word associations. This is especially relevant for weather-related texts, where sentiment is often implied through tone, metaphor, or contrast. As such, SentiBERT offers a strong foundation for developing robust classifiers in social media-based climate monitoring applications.

3.2.2. Architecture

SentiBERT, a model designed to understand compositional sentiment by leveraging the constituency structure of sentences. SentiBERT is composed of three key components: (1) the BERT model; (2) a semantic composition module with an attention mechanism; and (3) sentiment prediction components for both phrases and entire sentences. These three components are shown in Fig. 3, and a summary is provided below.

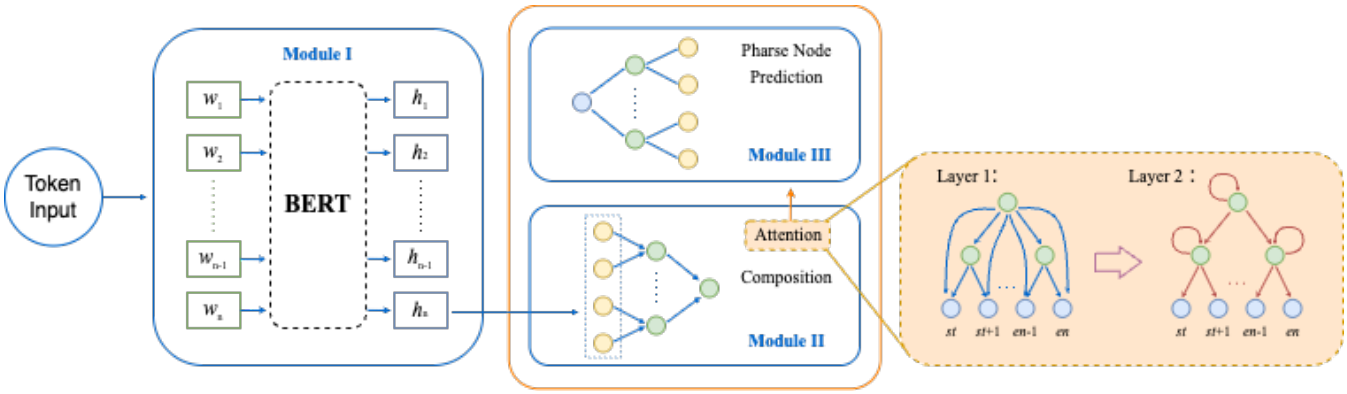

Fig. 3: Illustrates SentiBERT architecture. Module I is the BERT encoder; Module II is the semantic composition module built on attention; and Module III predicts phrase-level sentiment.

Based on Fig. 3, The semantic composition module employs two-layer attention network, the first layer (Attention to Tokens) constructs each phrase representation from its underlying tokens, and the second layer (Attention to Children) enhances those representations by integrating information from their sub-phrase nodes. SentiBERT integrates BERT (Devlin et al., 2019) as the core encoder to produce contextualized embeddings for every token in the input sentence. Semantic Composition Module of SentiBERT builds on those embeddings and the sentence constituency parse tree to yield rich phrase-level representations. To inject structural cues, SentiBERT employ a two-stage attention scheme: (1) Token-Level Attention – focuses on the relevant contextual token embeddings, and (2) Child-Node Attention – aggregates information from a phrase constituent sub-phrases. In combination, these layers refine each phrase representation in line with both its context. SentiBERT phrase-level sentiment predictor is trained using sentiment labels assigned to each phrase. SentiBERT uses cross-entropy as the loss function for learning the sentiment predictor. Motivated by BERT, SentiBERT training objective is composed of two parts: (1) Masked Language Modeling: Selected tokens are masked and the model is trained to predict them, enabling it to capture contextual information in the same method as BERT. (2) Phrase-Node Prediction: Using the phrase representations generated by the semantic composition module, the model is further trained to predict sentiment labels at each phrase node, thereby the model will learn compositional sentiment semantics. SentiBERT is implemented based on the HuggingFace library and initialized using pre-trained BERT-base and RoBERTa-base models, with a maximum sequence length of 128, 12 layers, and an embedding dimension of 768. For training on the SST-phrase dataset, the learning rate is set to 2×10^{-5}, the batch size to 32, and the number of training epochs to 3. To emphasize sentiment modeling in the masking mechanism, opinion words identified using SentiWordNet (Baccianella et al., 2010) are masked with a probability of 20%, while other words are masked with a probability of 15%. During fine-tuning

on downstream tasks, the learning rate ranges from 1×10^{-5} to 1×10^{-4}, the batch size is set to either 16 or 32, and the number of training epochs varies from 1 to 5. The Stanford CoreNLP API (Manning et al., 2014) is used to generate binary constituency trees for task-related sentences to maintain consistency with the SST-phrase setup. For sentence-level sentiment and emotion classification tasks, the objective is to correctly classify the root node of the tree structure, rather than relying on the [CLS] token representation as used in the original BERT.

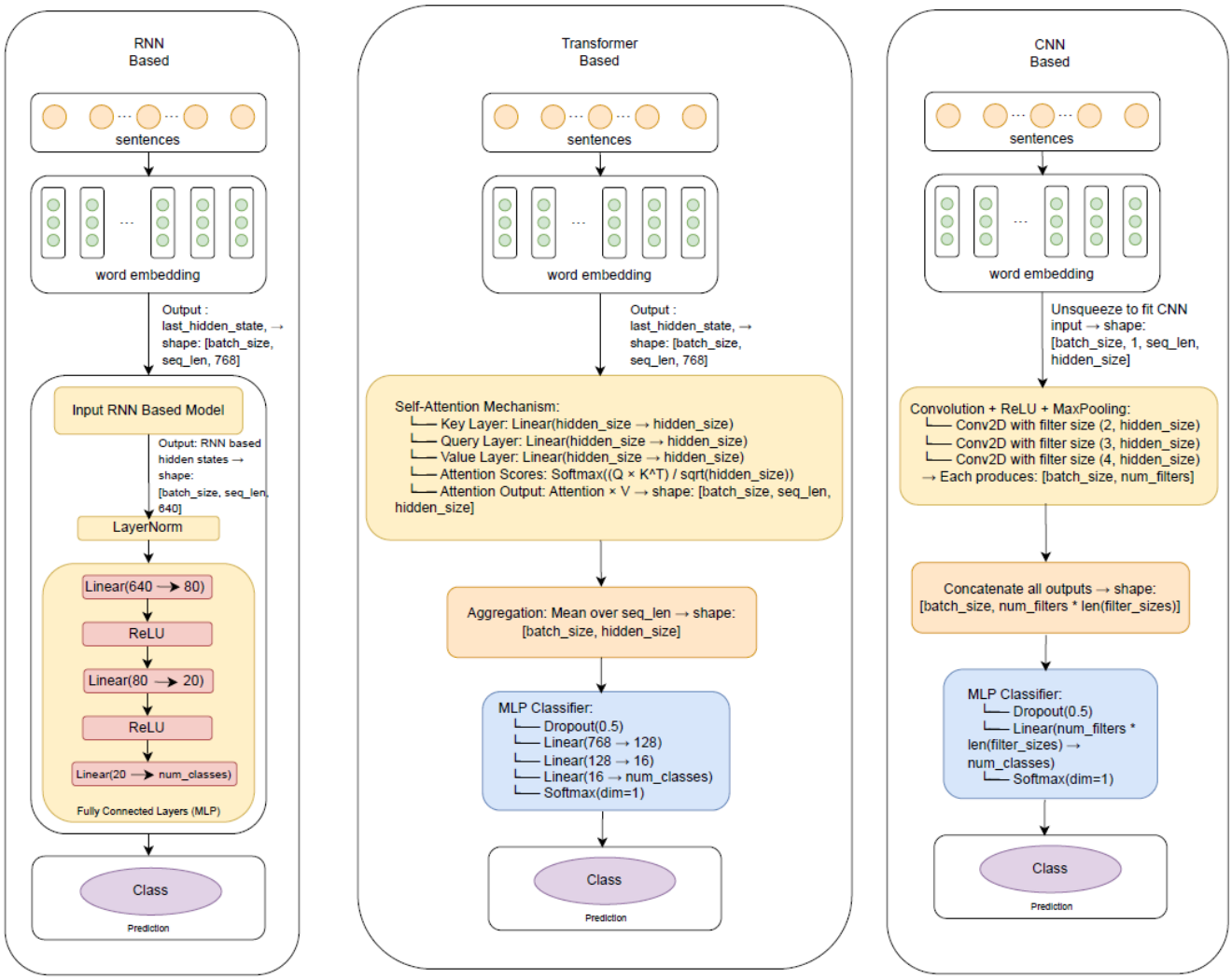

Fig. 4: Model architecture: RNN-based, Transformer-based, and CNN-based.

The RNN-based model uses layer followed by 3-layer Multi Layer Perceptron (MLP) with linear layers (640 → 80 → 20 → num_classes) and ReLU activation functions. The Transformer-based model leverages a self-attention mechanism by computing scores between Query, Key, and Value vectors. The resulting attention outputs are aggregated using mean pooling across the sequence and passed through MLP consisting of linear and dropout layers.The CNN-based model applies three Conv2D layers with different filter sizes (2, 3, and 4), followed by ReLU activation and max pooling. The outputs are concatenated and then passed through MLP for classification. All three models generate class predictions through a final Softmax layer. The architectures of these models are illustrated in Fig. 4

3.3. Classifier Model

3.3.1. Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU), a type of recurrent neural network (RNN), was introduced by Cho et al. in 2014. The GRU model mainly consists of an update gate and a reset gate. The update gate (u_t) assesses the

importance of new information, while the reset gate (r_t) helps retain memory by removing unnecessary long-term data. Additionally, the GRU incorporates a new candidate state $h_{\sim}$ and an updated hidden state $h_{\{t-1\}}$. The cyclic unit structure of the GRU is depicted in Fig. 5.

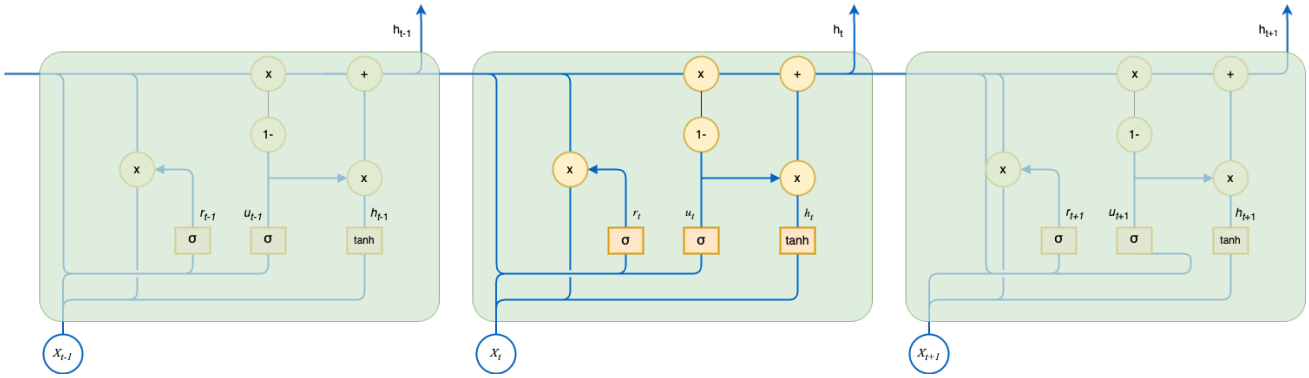

Fig. 5: GRU architecture.

The equations are as follows:

$$u_t = \sigma\left(W_u\left[h_{\{t-1\}}, x_t\right] + b_u\right) \tag{1}$$

$$r_t = \sigma\left(W_r\left[h_{\{t-1\}}, x_t\right] + b_r\right) \tag{2}$$

$$h_{\tilde{t}} = tanh\left(W_h\left[r_t * h_{\{t-1\}}, x_t\right] + b_h\right) \tag{3}$$

$$h_t = (1 - u_t) * h_{\{t-1\}} + u_t * h_{\tilde{t}} \tag{4}$$

Here, x_t represents the current GRU input value, W_u, W_r, and W_h are the weight matrices, $h_{\{t-1\}}$ denotes the output value from the previous timestep, b_u, b_r, and b_h are biases, and σ and $tanh$ refer to activation functions. This GRU configuration was selected to provide a balance between computational efficiency and the ability to model temporal dependencies in short, informal tweets. The bidirectional setup (Bi-GRU) allows the model to capture both past and future context, which is critical for disambiguating phrases such as “mendung banget, kayaknya mau hujan” that depend heavily on word order and temporal clues. By combining this optimized GRU setup with rich contextual embeddings from SentiBERT, the model effectively captures both sequential structure and sentiment nuances present in weather-related social media texts.

3.3.2. Bidirectional GRU (Bi-GRU)

The Bi-GRU model is a type of Recurrent Neural Network (RNN) that adeptly addresses some of the inherent limitations of traditional RNNs, particularly their struggle with long-term dependencies and unidirectional information flow. Unlike standard RNNs that process data in a single direction, the Bi-GRU architecture ingeniously utilizes two separate Gated Recurrent Unit (GRU) layers. One GRU layer is dedicated to handling data in the forward direction, meticulously following the chronological order of the time series or sequence. This forward pass allows the model to effectively capture dependencies and patterns that emerge from past observations, building a sequential understanding of the data as it unfolds. Concurrently, a second, independent GRU layer processes the exact same data in the reverse direction. This backward pass is crucial because it enables the model to incorporate vital information from future points in the sequence. By considering both preceding and succeeding contexts, the Bi-GRU gains a much richer and more complete understanding of the relationships within the data. This bidirectional setup is a paramount advantage, as it allows the Bi-GRU to gather comprehensive contextual information from both past and future points in the sequence, significantly improving its ability to recognize and interpret intricate

patterns, discern subtle nuances, and effectively capture long-range dependencies that might be missed by a purely unidirectional model.

In the context of this research, where weather-related text classification from social media involves language that is often noisy, ambiguous, and carries implicit sentiment, Bi-GRU inherent ability to model sequential dependencies and comprehend context from both directions becomes critically important. This model was chosen due to its superior capability in understanding contextual representations and sentiment nuances within short and noisy texts, a challenge often unaddressed by traditional machine learning models such as Support Vector Machines (SVM). Specifically, within the "Enhanced Bi-GRU" architecture proposed in this study, the Bi-GRU component, when combined with contextual embeddings from SentiBERT, is further able to significantly improve the understanding of implicit and complex sentiments in weather data sourced from Twitter. This synergistic approach leverages Bi-GRU strengths in sequence processing to effectively interpret the complex linguistic patterns prevalent in social media, thereby enhancing classification accuracy. The elegant and powerful Bi-GRU architecture is illustrated in detail in Fig. 6, visually representing how information flows and is processed in both directions.

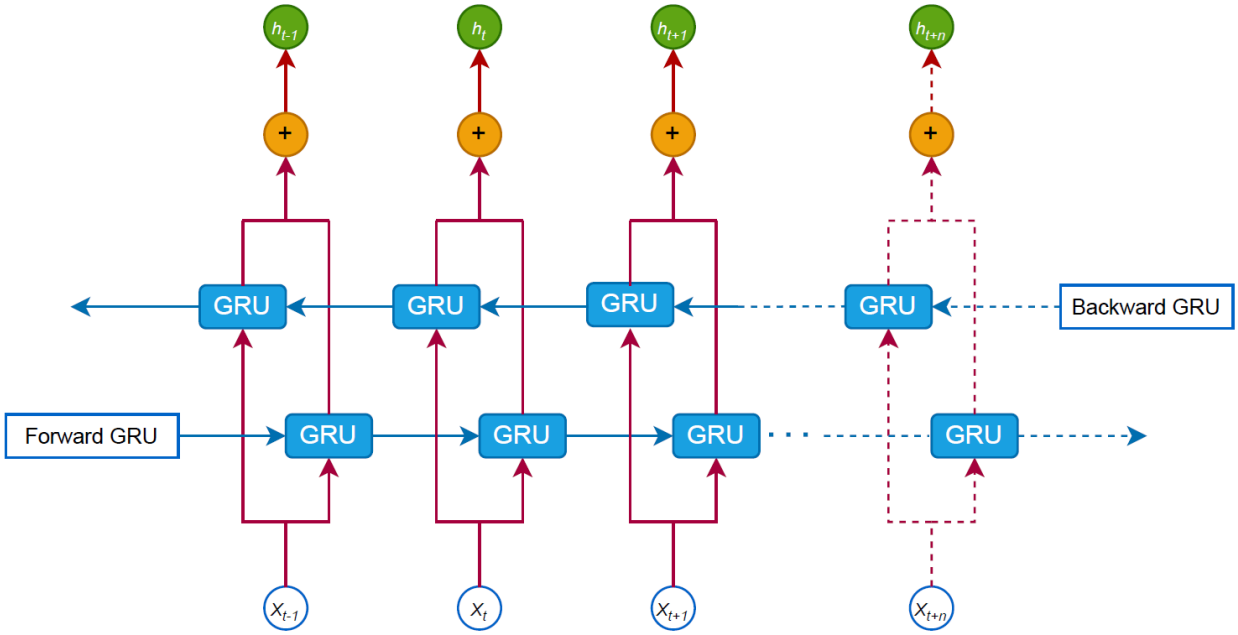

Fig. 6: Bi-GRU architecture.

3.3.3. SentiBERT with Enhanced Bi-GRU

The implemented Attention class represents a simple form of self-attention, often referred to as additive or soft attention, which aims to extract the most relevant information from a sequence of vectors by assigning different importance weights to each timestep. The process begins by applying a linear transformation to each feature vector in the input sequence x, which has the shape (batch_size, seq_len, hidden_size). This transformation maps each vector to a scalar attention logit, resulting in a tensor of shape (batch_size, seq_len, 1). These logits are then passed through a softmax function across the sequence length dimension to produce normalized attention weights, probability scores that indicate the relative importance of each timestep in the sequence. After that, each vector in the sequence is scaled (element-wise multiplied) by its corresponding attention weight. The weighted vectors are then summed along the time axis to produce a single context vector of shape (batch_size, hidden_size) for each instance in the batch. This context vector serves as a summary of the sequence, emphasizing the most informative parts, and is particularly useful for downstream tasks such as classification.

The architecture of the SentiBERT with Enhanced Bi-GRU model show in Algorithm 1, integrates contextual embeddings from SentiBERT with bidirectional sequence processing using Bi-GRU, and is further enhanced by a

self-attention mechanism and LeakyReLU activation (α = 0.11). The process begins with extracting features from SentiBERT, followed by Bi-GRU to capture forward and backward dependencies in the text. The output is then passed through a self-attention layer to emphasize the most relevant parts of the sequence. The resulting context vector is normalized and forwarded through fully connected layers equipped with dropout and LeakyReLU activation to maintain gradient flow and prevent overfitting. This combination enables the model to effectively classify weather-related tweets with high accuracy, especially when dealing with ambiguous or sentiment-rich language.

Algorithm 1 SentiBERT with Enhanced Bi-GRU Model (with Self-Attention)

Initialize base model with SentiBERT pre-trained weights
Define hyperparameters: number of classes ($num\_classes$), input size ($input\_size$), hidden size ($hidden\_size = 320$)
Create bidirectional GRU layer with $input\_size$, $hidden\_size$, 1 layer, and $batch\_first = True$
Initialize attention mechanism:
- Create a linear layer mapping from $hidden\_size \times 2$ to 1

Set up normalization layer (e.g., LayerNorm) for $hidden\_size \times 2$
Construct fully connected layers:
- Dropout layer with probability 0.5
- Linear layer from $hidden\_size \times 2$ to 80
- LeakyReLU activation with slope 0.11
- Linear layer from 80 to 20
- LeakyReLU activation with slope 0.11
- Linear layer from 20 to $num\_classes$

Enable gradient computation for all base model parameters
Forward Pass:
Extract raw outputs (last hidden state) from SentiBERT base model using input data
Pass tokens to Bi-GRU layer to get gru_output and discard hidden state
Apply attention mechanism to gru_output:
- Compute attention weights using linear layer and softmax along dimension 1
- Calculate weighted sum by multiplying gru_output with attention weights and summing along dimension 1

Normalize the attention output using the normalization layer
Feed the normalized output through fully connected layers to get final $outputs$
Return $outputs$ as the model prediction

3.4. Metric Evaluation

The performance of the proposed models was evaluated using four standard classification metrics: Overall Accuracy, Macro Precision, Macro Recall, and Macro F1 Score. These metrics are mathematically defined as follows:

Overall Accuracy: Measures the overall correctness of the model across all classes.

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Macro Precision: The average precision across all classes, where precision for each class is the ratio of correctly predicted positive instances to the total predicted positive instances for that class.

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{6}$$

$$Macro\ Precision = \frac{\sum_{c=1}^{n} Precision_c}{n} \tag{7}$$

Macro Recall: The average recall across all classes, where recall for each class is the ratio of correctly predicted positive instances to the total actual positive instances for that class.

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{8}$$

$$Macro\ Recall = \frac{\sum_{c=1}^{n} Recall_c}{n} \tag{9}$$

Macro F1 Score: The average F1 score across all classes, where the F1 score for each class is the harmonic mean of precision and recall for that class.

$$F1\ score = 2 * \frac{Precision_c * Recall_c}{Precision_c + Recall_c} \tag{10}$$

$$Macro\ F1\ Score = \frac{\sum_{c=1}^{n} F1\_c}{n} \tag{11}$$

where:

- TP (True Positives) is the number of correctly predicted positive instances.
- TN (True Negatives) is the number of correctly predicted negative instances.
- FP (False Positives) is the number of incorrectly predicted positive instances.
- FN (False Negatives) is the number of incorrectly predicted negative instances.
- For a specific class c, TP_c, FP_c, and FN_c represent the True Positives, False Positives, and False Negatives for that class.
- n is the number of classes.

4. Result and Analysis

This section presents the results of weather-related tweet classification using a hybrid model that combines BERT-based embeddings with an Enhanced Bi-GRU classifier. We experimented with several pretrained models BERT, DistilBERT, RoBERTa, and SentiBERT, paired with RNN, CNN, and Transformer-based classifiers. BERT was chosen as the tokenizer for its superior ability to generate context-aware embeddings, as also demonstrated by Kumawat et al. (2024) in disaster tweet classification using the Hugging Face AutoTokenizer.

4.1. Experiment Analysis

All experiments were implemented using PyTorch 2.4.0 and conducted on an NVIDIA A6000 GPU with 48 GB of memory. The classification task employed five weather categories: sunny, cloudy, rainy, extreme, and other, evaluated using Overall Accuracy, Macro Precision, Macro Recall, and Macro F1 Score. Macro metrics were selected as they ensure equal contribution from each class regardless of class imbalance. The training process used 5-fold stratified cross-validation with an 80:20 train–test split in each fold, without a separate validation set. We combined the BERT model with several deep learning classifiers, and the Bi-GRU classifier consistently outperformed the others when paired with both BERT and DistilBERT tokenizers. Performance was assessed using the four standard classification metrics, as previously described.

Table 1: Results of first scheme of experiments.

Embedding & Tokenizer	**Classifier**	**Overall Acc ↑**	**Macro Prec ↑**	**Macro Recall ↑**	**Macro F1 ↑**	**Cost time (s) ↓**
BERT	Transformer	81.98	82.48	81.52	81.62	**3.15**
BERT	BiLSTM	82.77	83.19	82.10	82.35	3.42
BERT	Bi-GRU	**84.43**	**85.32**	**84.09**	**84.41**	3.42
BERT	LSTM	84.11	84.76	83.67	84.01	3.25
BERT	GRU	82.72	83.4	82.05	82.44	3.26
BERT	CNN	75.07	76.18	73.33	73.54	4.14
BERT	RNN	82.93	83.75	82.44	82.84	4.39
DistilBERT	Transformer	83.59	84.55	83.19	83.52	**1.75**
DistilBERT	BiLSTM	82.19	82.64	81.63	81.89	2.03
DistilBERT	Bi-GRU	83.92	**84.88**	83.56	83.86	2.0
DistilBERT	LSTM	**84.11**	84.54	**83.89**	**84.05**	1.87
DistilBERT	GRU	82.56	83.23	82.3	82.4	1.85
DistilBERT	CNN	72.84	74.05	71.12	71.14	2.71
DistilBERT	RNN	82.68	83.39	82.13	82.46	1.79
RoBERTa	Transformer	81.75	82.21	81.41	**82.49**	**3.33**
RoBERTa	BiLSTM	81.79	81.92	**81.44**	81.33	3.64
RoBERTa	Bi-GRU	80.97	81.79	80.28	80.44	3.67
RoBERTa	LSTM	**81.96**	**82.34**	81.32	81.7	3.47
RoBERTa	GRU	81.27	82.19	81.08	81.3	3.46
RoBERTa	CNN	64.16	66.46	64.84	64.61	4.39
RoBERTa	RNN	79.53	79.81	78.98	79.21	3.97
SentiBERT	Transformer	86.86	87.04	86.91	86.92	**2.76**
SentiBERT	BiLSTM	86.06	86.68	86.05	86.23	2.93
SentiBERT	Bi-GRU	**86.99**	**87.66**	**86.98**	**87.22**	4.01
SentiBERT	LSTM	86.12	86.33	86.26	86.26	2.80
SentiBERT	GRU	86.83	87.45	86.84	86.98	2.83
SentiBERT	CNN	75.05	77.49	73.4	73.85	3.48
SentiBERT	RNN	86.36	87.1	86.39	86.62	2.79
SentiBERT	Bi-GRU	86.99	87.66	86.98	87.22	4.01
SentiBERT	Bi-GRU + Self-Attention	87.75	87.89	87.95	87.83	2.92
SentiBERT	Bi-GRU + Self-Attention + LeakyReLU (α = 0.10)	87.89	88.33	87.94	88.08	2.94
SentiBERT	**Bi-GRU + Self-Attention + LeakyReLU (α = 0.11)**	**88.03**	**88.49**	**88.12**	**88.25**	2.93
SentiBERT	Bi-GRU + Self-Attention + LeakyReLU (α = 0.12)	87.54	88.03	87.65	87.78	**2.91**
SentiBERT	Bi-GRU + Self-Attention + LeakyReLU (α = 0.09)	87.78	88.24	87.88	88.00	2.95

The combination of the BERT tokenizer with the Bi-GRU classifier based on Table 1 shows the best performance across all key evaluation metrics, achieving the highest values in Overall Accuracy (84.43%), Macro Precision (85.32%), Macro Recall (84.09%), and Macro F1 (84.41%). This indicates that Bi-GRU is more effective at capturing sequential information compared to the other classifiers. Although the Transformer-based model has the fastest computation time (3.15 seconds), its performance is still lower than that of Bi-GRU. Overall, these results confirm that Bi-GRU is the most optimal classifier to be used with the BERT tokenizer in this experimental scheme.

The results show based on Table 1 that the Bi-GRU classifier, when combined with the DistilBERT tokenizer, achieved the highest performance across all metrics, including Overall Accuracy (83.92%) and Macro F1 (83.86%). Although the Transformer model had the shortest computation time (1.75s), its accuracy and F1 score were slightly lower. LSTM also delivered strong results, however still fell short compared to Bi-GRU. These findings confirm that Bi-GRU is the most effective classifier in the DistilBERT-based setup.

The combination of SentiBERT with the Bi-GRU classifier achieved the best performance, recording the highest values in Overall Accuracy (86.99%), Macro Precision (87.66%), Macro Recall (86.98%), and Macro F1 (87.22%). This indicates that Bi-GRU effectively captures feature representations compared to other classifiers in this context. Although the Transformer model had the fastest computation time (2.76 seconds), its classification performance was still below that of Bi-GRU. Overall, Bi-GRU is the most superior classifier when paired with the SentiBERT tokenizer in this experimental scheme.

As shown in Table 1, the Transformer model, despite being the most recent architecture, did not demonstrate the best performance compared to earlier models. One of the main reasons is its parallel processing mechanism, which requires high-quality input data with minimal noise to function optimally. In this study, the preprocessing steps were limited to lowercasing and stopword removal, which likely left a significant amount of noise in the data, thereby negatively affecting the model's performance. This condition led the Transformer to overfit the existing noise. Therefore, we recommend that future research consider applying more comprehensive preprocessing techniques, such as lemmatization and noise filtering, to improve data quality and enhance the overall performance of the Transformer model.

This experiment demonstrates that different tokenizer-classifier combinations vary in performance across five key metrics. The SentiBERT with Bi-GRU combination stands out as the best, achieving an overall accuracy of 86.99% and macro F1 score of 87.22%, despite having a relatively higher processing time of 4.01 seconds. For applications requiring faster inference, DistilBERT with Transformer offers the quickest time at 1.75 seconds with decent accuracy of 83.59%. Overall, Bi-GRU proves to be the most reliable and powerful classifier across almost all tokenizers, while CNN performs the weakest in this text classification task. Therefore, it is recommended to use SentiBERT + Bi-GRU for the highest accuracy and DistilBERT + Transformer when processing speed is a priority.

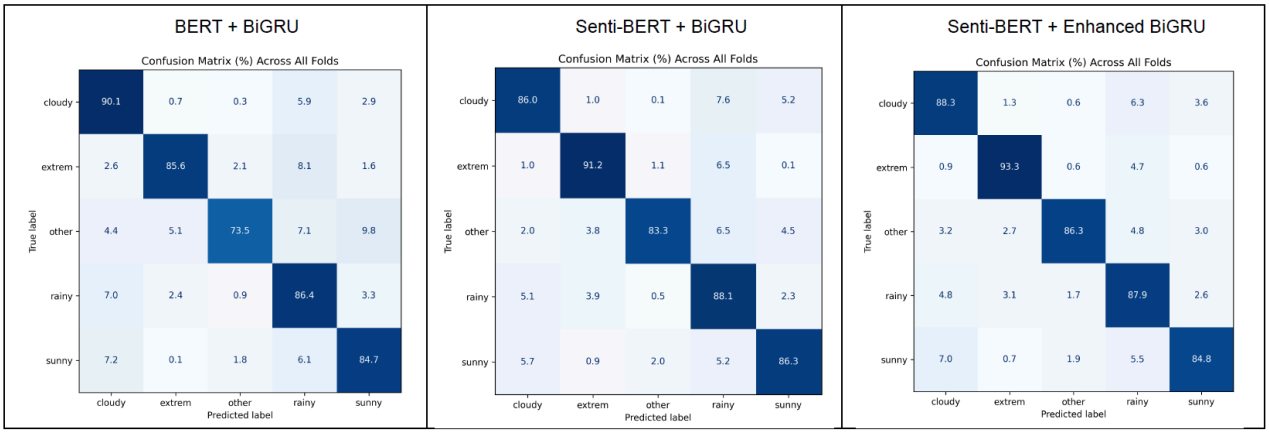

Fig. 7: Confusion matrix comparison of BERT+Bi-GRU, Senti-BERT+Bi-GRU, and Senti-BERT + Enhanced Bi-GRU.

Fig. 7 presents the confusion matrix comparing three models: BERT + Bi-GRU, SentiBERT + Bi-GRU, and the proposed SentiBERT + Enhanced Bi-GRU. The SentiBERT + Enhanced Bi-GRU model showed significant improvements in correctly classifying the extreme and other categories, which are generally more difficult due to data imbalance and vague semantics. The presence of the word “hujan” (rain) in multiple classes, as illustrated in the word cloud (Fig. 3), explains some of this complexity. Tweets mentioning “hujan” can refer to typical rain (rainy), potential rain (cloudy), or severe weather (extreme), depending on the context and tone. The Enhanced Bi-GRU effectively disambiguated these cases by focusing on contextual and sentiment-rich cues, as shown by the reduced misclassifications in the matrix.

Overall, the experimental results confirm that the integration of SentiBERT with Enhanced Bi-GRU provides a robust and context-aware solution for weather-related tweet classification. The hybrid model not only improves predictive accuracy across all categories but also demonstrates better handling of noisy, ambiguous, and sentiment-laden text data typically found in social media platforms.

4.2. Lexical Features Analysis

The purpose of lexical features analysis is to identify distinctive words or phrases that serve as key indicators in text classification, particularly in the context of weather categories. Its main goal is to understand how terms like “sunny”, “rain”, or “cloudy” are used in sentences and how their presence can influence the predictions made by classification models. This analysis is crucial for recognizing potential ambiguity in natural language that may lead to overlap between weather classes.

In the example sentences analyzed in Table 2, it is evident that a single sentence often contains lexical cues from multiple weather categories. For instance, a sentence that includes the words “cerah”, “mendung”, “hujan”, and “deras” may confuse the model, as it signals both the sunny and rainy classes. Therefore, this analysis helps highlight key challenges in natural language processing for multi-context classification tasks, such as text-based weather prediction.

Table 2: Example sentences for each weather class with potential lexical features.

Class	Example Sentence
other	1. “ibarat langit sore tidak selalu cerah bisa mendung bisa hujan dan derasnya tidak mampu kamu bendung terima kasih untuk segalanya”
	2. “arti ibarat mimpi tentang hujan mimpi hujan deras hujan gerimis kehujanan hingga berteduh dari hujan”
sunny	1. “dari pagi langitnya cerah banget terus tadi kayaknya sempat hujan sebentar terus sore ini cerah lagi hihi”
	2. “pagi mendung siang panas”
cloudy	1. “di tempat ku mendung mengundang gerimis kak dari pagi”
	2. “mendung paling entar sore hujan”
rainy	1. “tidak apa apa hujan nya sekarang yang penting nanti siang cerah”
	2. “cuaca lagi enggak tentu siang panas banget terus sore langsung hujan bagaimana enggak pada tumbang”
extrem	1. “sudah tanda kegelapan datang di langit setiap jelang sore angin berbau hujan berhembus kencang”
	2. “waspada potensi hujan disertai kilat petir dan angin kencang di beberapa wilayah di maluku utara”

4.2.1. Ambiguity Analysis Across Weather Categories

In text-based weather classification, ambiguity arises when a single sentence contains words or phrases that represent more than one weather class category. This poses a major challenge in natural language processing, as the model may receive conflicting signals from a single input.

- **Other Class:** “arti ibarat mimpi tentang hujan mimpi hujan deras hujan gerimis kehujanan hingga berteduh dari hujan”.
  Lexical Features: “hujan gerimis” → **rainy**, “hujan deras” → **extrem**

- **Sunny Class:** “dari pagi langitnya cerah banget terus tadi kayaknya sempat hujan”.
  Lexical Features: “cerah banget” → **sunny**, “sempat hujan” → **rainy**
- **Cloudy Class:** “pagi mendung siang panas”.
  Lexical Features: “mendung” → **cloudy**, “panas” → **sunny**
- **Rainy Class:** “mendung paling entar sore hujan”.
  Lexical Features: “mendung” → **cloudy**, “hujan” → **rainy**
- **Extrem Class:** “cuaca lagi enggak tentu siang panas banget terus sore langsung hujan pohon tumbang”.
  Lexical Features: “panas banget” → **sunny**, “langsung hujan” → **rainy**, “pohon tumbang” → **extrem**

SentiBERT is used to help categorize weather-related sentences based on the emotional tone conveyed in the text (positive, neutral, negative). This sentiment-based separation assists in the initial filtering process to map sentences to the most appropriate weather class, especially when lexical ambiguity is present, as illustrated in Fig. 8. In addition, based on TF-IDF analysis, several frequently occurring keywords reinforce this sentiment-weather mapping. Words like “cerah” and “panas” are dominant in positive-sentiment sentences; “mendung”, “prakiraan”, and “ibarat” frequently appear in neutral contexts; while “hujan”, “deras”, and “angin” stand out in negative-sentiment expressions. These TF-IDF-derived features further support the classification process by highlighting the most informative lexical items in each sentiment group.

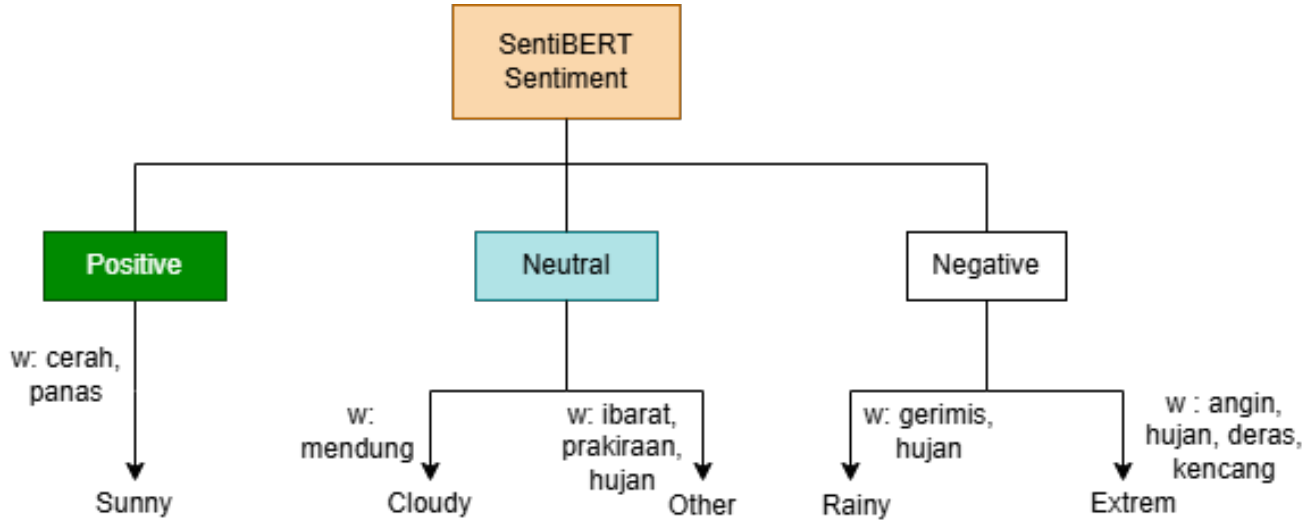

Fig. 8: SentiBERT sentiment flow categories on weather class.

Positive sentiment is typically associated with the Sunny weather class. Common lexical indicators include words such as “cerah” (clear) and “panas” (hot), which convey a pleasant or comfortable atmosphere. These expressions help the model recognize that the sentence reflects favorable weather conditions. One key advantage of this sentiment-based cue is its ability to prevent misclassification of sentences that contain ambiguous terms. For example, a phrase like “tidak hujan” (not raining) or “akhirnya cerah” (finally clear) may still express positivity despite mentioning rain-related terms. By relying on the overall sentiment, the model can more accurately assign such sentences to the Sunny class.

Neutral sentiment typically corresponds to the Cloudy and Other weather classes. Key lexical features include words such as “mendung” (cloudy), “ibarat” (like/as), “prakiraan” (forecast), and “hujan” (rain), which are often used in informative or emotionless statements. These sentences usually lack strong emotional cues and instead focus on describing conditions or making predictions. This sentiment category is particularly useful for distinguishing between literal weather forecasts and metaphorical or non-literal expressions, especially those containing rain-related words that do not indicate actual weather, as is often the case in the Other class.

Negative sentiment is generally associated with the Rainy and Extreme weather classes. Typical keywords for the Rainy class include “gerimis” (drizzle) and “hujan” (rain), while the Extreme class is characterized by terms like “angin” (wind), “deras” (heavy), and “kencang” (strong). These expressions often appear in sentences that convey warnings, dangers, or descriptions of severe weather conditions, making them inherently negative in tone. This sentiment categorization is especially beneficial for distinguishing between light rain events and hazardous situations. By analyzing the intensity and context of the vocabulary used, the model can better differentiate between the Rainy and Extreme classes, improving classification accuracy.

4.2.2. Analyze based on Tree Structure

This section presents a compositional sentiment analysis using SentiBERT, visualized through hierarchical tree structures. By breaking down weather-related sentences into sentiment layers (positive, neutral, negative), the model can better interpret lexical ambiguities and improve classification accuracy across weather categories.

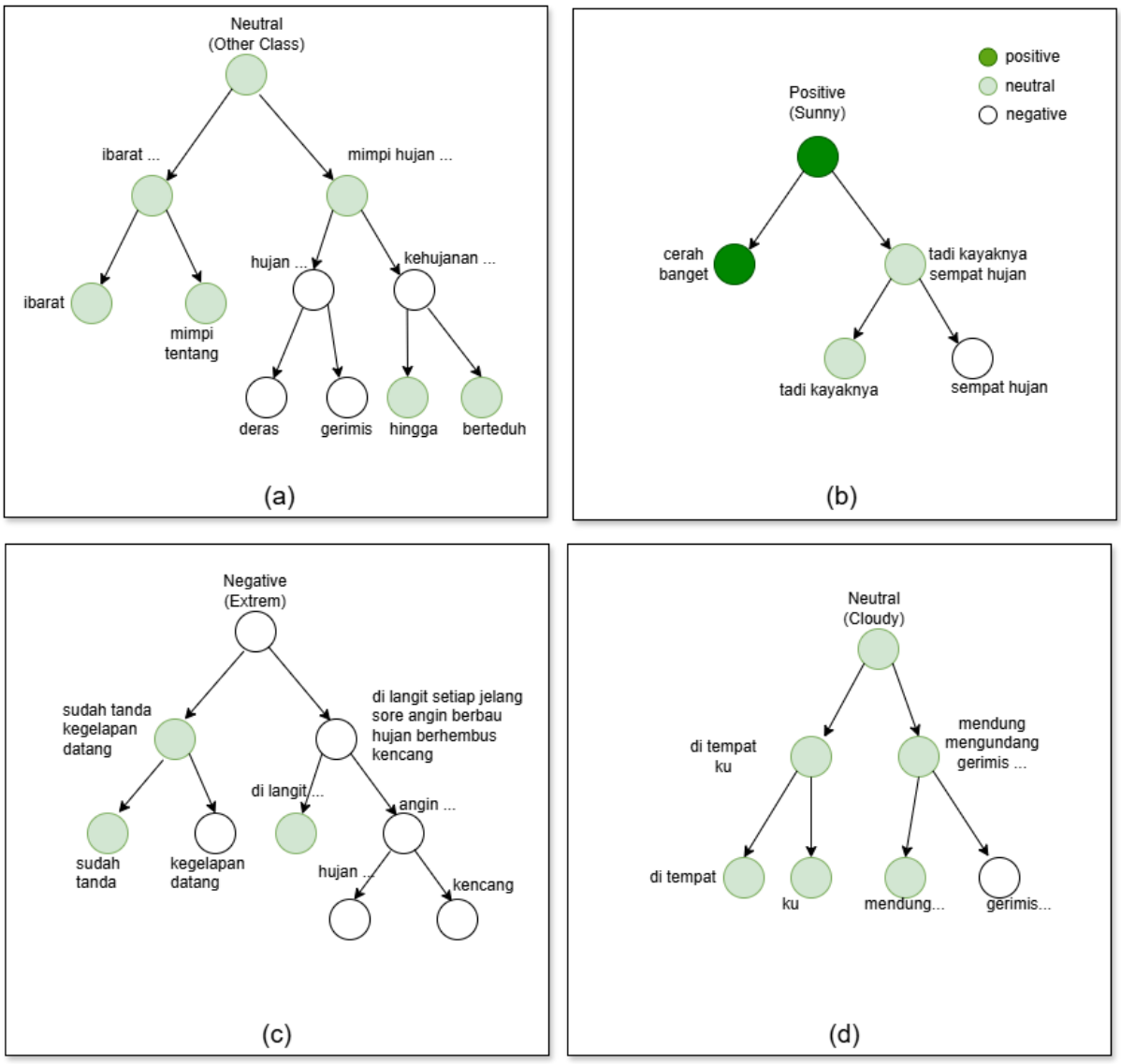

Fig. 9: Examples illustrating the interpretability of compositional sentiment analysis using SentiBERT in weather classification.

Based on Fig. 9, we present an analysis using a tree structure, illustrating several example sentences for each weather class category as follows.

- **Other Class:** “arti ibarat mimpi tentang hujan mimpi hujan deras hujan gerimis kehujanan hingga berteduh dari hujan”.

Tree Analysis: Based on Fig. 9 (a), The sentence in this tree consists of non-literal phrases such as “ibarat” (like/as if), “mimpi tentang” (dream about), “mimpi hujan” (dream of rain), and “kehujanan hingga berteduh” (getting caught in the rain until taking shelter), all of which are categorized as neutral sentiment and mapped to the “Other” weather class. At the root level, the sentence is considered neutral because it does not directly describe actual weather conditions, but instead conveys metaphorical or imaginative language.

The first subtree stems from the phrase “ibarat mimpi tentang”, in which: The words “ibarat” and “mimpi tentang” are considered neutral as they express a comparison or analogy rather than a real event. The second subtree contains the phrase “mimpi hujan kehujanan hingga berteduh”:

“Mimpi hujan” is still categorized as neutral, but when followed by words such as “hujan deras” (heavy rain) and “gerimis” (drizzle), a negative nuance emerges, as these terms denote unpleasant weather conditions.

“Kehujanan” (getting rained on) is also considered negative due to its root word “hujan” (rain), yet the following phrase “hingga berteduh” (until taking shelter) returns the tone to neutral, as it reflects a resolving or coping action.

Overall, despite the presence of words with negative connotations, the context of dreaming and the metaphorical style lead the sentence to remain classified as neutral.

- **Sunny Class:** “dari pagi langitnya cerah banget tadi kayaknya sempat hujan”.
  Tree Analysis: Based on Fig. 9 (b), the sentence has been broken down into compositional sentiment nodes. At the root, the overall sentiment is labeled positive, aligning with the Sunny class due to the strong presence of the phrase “cerah banget” (very sunny), which clearly indicates pleasant weather conditions. However, as the tree branches out, lexical ambiguity emerges. The phrase “tadi kayaknya sempat hujan” is decomposed further:

  “tadi kayaknya” is interpreted as neutral, reflecting uncertainty or speculation.

  “sempat hujan” introduces a negative sentiment because it denotes an actual rain event.

  This analysis shows that although the dominant sentiment is positive, the inclusion of speculative or contradictory weather terms leads to multi-sentiment blending, making it challenging for models to classify it purely as Sunny or Rainy. The use of SentiBERT and tree-based interpretability helps to expose this lexical complexity and improve classification accuracy.
- **Extrem Class:** “sudah tanda kegelapan datang di langit setiap jelang sore angin berbau hujan berhembus kencang”.
  Tree Analysis: Based on Fig. 9 (c), the sentence is divided into two main branches. The left branch, “sudah tanda kegelapan datang” (already a sign that darkness is coming), is broken down into “sudah tanda” (already a sign) and “kegelapan datang” (darkness is coming), both of which semantically indicate an early warning of an approaching extreme weather condition, metaphorically represented by the coming darkness. The right branch, “di langit setiap jelang sore angin berbau hujan berhembus kencang” (in the sky every late afternoon, the wind smells like rain and blows strongly), begins with a spatial cue “di langit…” (in the sky…), then elaborates on weather elements such as “hujan” (rain) and “angin” (wind), with further details like “berhembus” (blows) and “kencang” (strong). The phrase “angin berbau hujan” (wind smells like rain) signals atmospheric changes that anticipate rainfall, while “berhembus kencang” (blows strongly) reinforces the notion of extreme weather.
- **Cloudy Class:** “mendung paling entar sore hujan”.
  Lexical Features: Based on Fig. 9 (d), the combination of “di tempatku” (“at my place”) and “mendung mengundang gerimis” (“cloudy skies inviting drizzle”) indicates a specific location and describes a cloudy weather condition with the possibility of light rain. There is no strong emotional expression present, nor is there any indication of extreme weather. Therefore, the overall sentiment is considered neutral.

SentiBERT with a tree decomposition approach has proven effective in capturing contextual patterns in descriptive weather-related sentences. By breaking down sentences into smaller semantic units, the model can isolate key phrases such as “cerah banget”, “angin berhembus kencang,” or “mendung mengundang gerimis”

allowing for more accurate sentiment and weather intensity assessments. This approach also enables SentiBERT to distinguish between relevant and irrelevant contexts, such as separating location references “di tempat ku” from actual weather descriptions. Even in metaphorical or ambiguous cases, like the phrases “mimpi tentang hujan” or “ibarat kehujanan” the model is still able to classify them appropriately as neutral. Furthermore, interactions between phrases are thoroughly considered, for example, a strongly positive phrase like “very sunny” can dominate the overall interpretation of a sentence, even if it includes elements of rain. Thus, tree-based decomposition not only enhances SentiBERT’s semantic understanding but also makes it more sensitive to linguistic structure and nuance in the context of text-based weather classification.

5. Conclusion

This study successfully addressed the challenges inherent in classifying weather-related textual data that are often characterized by noise, ambiguity, and subtle emotional expressions, particularly within the informal linguistic context of social media platforms such as Twitter. The proposed hybrid architecture, which integrates SentiBERT, a sentiment-aware transformer model, with an Enhanced BiGRU network equipped with self-attention and LeakyReLU activation, demonstrates superior performance in capturing both contextual representations and sequential dependencies. Empirical evaluations using an Indonesian-language dataset encompassing five weather categories, sunny, cloudy, rainy, extreme, and other, indicated that the model achieved an overall accuracy of 88.03% and a macro F1-score of 88.25%, outperforming all baseline models. These results confirm that integrating sentiment-specific embeddings and sequential modeling significantly enhances classification accuracy for weather-related text data.

Nonetheless, this study is not without limitations. The preprocessing procedures implemented were limited to lowercase conversion and stopword removal, without incorporating more advanced normalization methods such as lemmatization, slang expansion, or emoji-to-text translation. Consequently, residual noise from informal language patterns, typographical variations (e.g., elongated words such as “hujaaan”), and mixed-language expressions may have introduced distortions in the embedding space generated by SentiBERT. Furthermore, social media-specific features, including hashtags, emoticons, and user mentions, were excluded from the modeling process. These elements often convey implicit affective or contextual information (e.g., sarcasm or emotional reinforcement through emojis) that can influence sentiment interpretation. The omission of these features potentially limited the model’s ability to fully capture the semantic richness of social media discourse.

Another limitation concerns the dataset scale and representativeness. The corpus used in this study consisted of 797 tweets, which, although sufficient for model prototyping, may not fully reflect the lexical and contextual diversity of broader weather-related discourse across regions and time frames. Moreover, the inclusion of the “other” category, while enhancing robustness against ambiguous cases, may have introduced overlapping semantics between adjacent classes, thereby moderating classification precision.

Future research is encouraged to address these limitations by incorporating more sophisticated preprocessing pipelines that can handle slang normalization, emoji sentiment mapping, and multilingual code-switching. Expanding the dataset to include a larger and more balanced distribution of weather-related tweets would also enhance generalizability. Additionally, integrating multimodal information, such as images, videos, or geospatial data, alongside advanced domain-adaptive pretraining using models like ClimateBERT fine-tuned for the Indonesian context, represents a promising direction for improving real-time weather monitoring and disaster early warning applications.

CRediT authorship contribution statement

M. A. Syaefudin: Conceptualization, Methodology, Software, Investigation, Resources, Data Curation, Writing – Original Draft, Visualization, Funding Acquisition. **A. I. Jati:** Resources, Writing – Review & Editing, Supervision. **H. Tsaniya:** Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision, Project Administration. **C. Fatichah:** Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision,

Project Administration. **D. Purwitasari:** Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision, Project Administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Declaration of Generative AI and AI-assisted Technologies in The Writing Process

The authors used generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and take full responsibility for the final publication.

References

[1] S. Jardim and C. Mora, "Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 199–206. doi: 10.1016/j.procs.2021.12.006.

[2] A. Natayu, F. Kamila, I. Dananjaya, R. Reflin, and M. Fikri, "Understanding the Climate Behavior Through Data Interpretation: Java-Bali-Nusa Tenggara Case," *Indonesian Journal of Computing, Engineering and Design (IJoCED)*, 2021, doi: 10.35806/ijoced.v3i2.184.

[3] S. Gaztelumendi, "Towards the operational use of tweets data in high impact weather scenarios: data mining and analytics in Basque Country.," 2021, doi: 10.5194/ems2021-245.

[4] S. Deb and A. Chanda, "Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data," *Machine Learning with Applications*, vol. 7, p. 100253, 2022, doi: 10.1016/j.mlwa.2022.100253.

[5] A. Tounsi and M. Temimi, "A systematic review of natural language processing applications for hydrometeorological hazards assessment," *Natural Hazards (Dordrecht, Netherlands)*, vol. 116, pp. 2819–2870, 2023, doi: 10.1007/s11069-023-05842-0.

[6] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," vol. 13. Association for Computing Machinery, 2022. doi: 10.1145/3495162.

[7] K. Purwandari, T. W. Cenggoro, J. Sigalingging, and B. Pardamean, "Twitter-based classification for integrated source data of weather observations," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, pp. 271–283, 2023, doi: 10.11591/ijai.v12.i1.pp271-283.

[8] K. Purwandari, J. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class Weather Forecasting from Twitter Using Machine Learning Aprroaches," *Procedia Computer Science*, vol. 179, pp. 47–54, 2021, doi: 10.1016/j.procs.2020.12.006.

[9] H. Otudi, S. Gupta, N. Albarakati, and Z. Obradovic, "Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis," *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023, doi: 10.1145/3625007.3627298.

[10] R. S. Wilkho, S. Chang, and N. G. Gharaibeh, "FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events," *Advanced Engineering Informatics*, vol. 59, p. 102293, 2024, doi: https://doi.org/10.1016/j.aei.2023.102293.

[11] S. Debata and S. Elango, "Empowering flood forecasting through meteorological and social media data," *International Journal of Information Technology*, vol. 16, no. 6, pp. 3757–3770, Aug. 2024, doi: 10.1007/s41870-024-01961-4.

[12] M. Wankhade, A. C. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.

[13] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research," *IEEE Transactions on Affective Computing*, vol. 14, pp. 108–132, 2020, doi: 10.1109/TAFFC.2020.3038167.

[14] V. S. Anoop, T. K. A. Krishnan, A. Daud, A. Banjar, and A. Bukhari, "Climate Change Sentiment Analysis Using Domain Specific Bidirectional Encoder Representations From Transformers," *IEEE Access*, vol. 12, no. , pp. 114912–114922, 2024, doi: 10.1109/ACCESS.2024.3441310.

[15] D. Adwaith, A. K. Abishake, S. V. Raghul, and E. Sivasankar, "Enhancing multimodal disaster tweet classification using state-of-the-art deep learning networks," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18483–18501, 2022, doi: 10.1007/s11042-022-12217-3.

[16] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences*, vol. 13, no. 6, 2023, doi: 10.3390/app13063915.

[17] D. Effrosynidis, A. I. Karasakalidis, G. Sylaios, and A. Arampatzis, "The climate change Twitter dataset," *Expert Systems with Applications*, vol. 204, p. 117541, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117541.

[18] L. Shi and D. Zhao, "Automatic Identification of Helpful Information on Social Media During Natural Disaster Based on Word2Vec and Bert," in *2023 18th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2023, pp. 234–239. doi: 10.1109/ISKE60036.2023.10481342.