BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 3501-3518
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp3501-3518

A HYBRID BERT-BILSTM-ATTENTION MODEL FOR PUBLIC SENTIMENT ANALYSIS ON A NATIONAL SOCIAL INSURANCE PROVIDER IN INDONESIA

Syaiful Anam [1*], **Nur Atiqah Sia Abdullah** [2], **Hilmi Aziz Bukhori** [3], **Avin Maulana** [4], **Regina Vincentia** [5]

[1,3,4,5]*Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya*
Jln. Veteran, Malang, 65145, Indonesia

[2]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA*
Jln. Ilmu 1/1, Shah Alam, Selangor, 40450, Malaysia

*Corresponding author's e-mail: * **syaiful@ub.ac.id***

Article Info

Article History:

Received: 07th November 2025
Revised: 21st April 2026
Accepted: 04th June 2026
Available online: 24th August 2026

Keywords:

Attention;
BERT;
BiLSTM;
Sentiment analysis;
Social insurance.

ABSTRACT

Sentiment analysis in Indonesian social media presents significant challenges due to informal language, code-mixing, and sentiment ambiguity in multi-clause expressions. While transformer-based models such as IndoBERT effectively capture contextual semantics, they may struggle to model sentiment transitions and resolve conflicting polarity in noisy and heterogeneous text. To address this limitation, this study employs a hybrid BERT–BiLSTM–Attention model that integrates contextual, sequential, and attention-based representations. Although the architecture itself is not novel, the contribution lies in its systematic integration and empirical evaluation under realistic conditions. Experimental results show that the proposed model achieves an accuracy of 0.85 and a Macro-F1 score of 0.85, outperforming the IndoBERT baseline (0.83) by approximately 2.4%. This improvement is statistically significant ($p = 0.027$) and supported by effect size analysis, indicating meaningful performance gains. Robustness evaluation under controlled perturbations—including slang injection, character elongation, emoji usage, code-mixing, and class imbalance (up to 30% minority downsampling)—shows only minor performance degradation (0.02–0.03 Macro-F1), demonstrating stable generalization under noisy conditions. These findings provide empirical evidence that the integration of sequential modeling and attention improves the handling of sentiment transitions and multi-clause structures beyond transformer-only approaches, offering practical value for real-world sentiment analysis in Indonesian social media.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/).

How to cite this article:

S. Anam, N. A. S. Abdullah, H. A. Bukhori, A. Maulana and R. Vincentia., "A HYBRID BERT-BILSTM-ATTENTION MODEL FOR PUBLIC SENTIMENT ANALYSIS ON A NATIONAL SOCIAL INSURANCE PROVIDER IN INDONESIA", *BAREKENG: J. Math. & App.*, vol. 20, no. 4, pp. 3501-3518, Dec, 2026.

Copyright © 2026 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Public sentiment analysis on social media has become increasingly important for evaluating institutional trust and service quality. In recent years, social media usage in Indonesia has experienced significant growth, reflecting a broader global trend. Recent reports indicate that social media penetration in Indonesia has increased significantly, rising from 59% of the population in 2020 to approximately 68.9% in early 2022, with average daily usage approaching four hours [1]. This rapid growth underscores social media's deep integration into Indonesian society, influencing communication, education, health initiatives, and commerce. Platforms such as YouTube and Instagram have further shaped marketing strategies and brand trust during the COVID-19 pandemic [2]. Consequently, social media provides both opportunities for engagement and challenges related to misinformation and public perception [3].

For public institutions, particularly social insurance providers, sentiment analysis not only captures reactions to policy changes and socioeconomic events [4] but also reflects public trust in government programs [5]. During crises such as the COVID-19 pandemic, it proved vital in addressing misinformation and strengthening public communication [6], [7]. Moreover, effective sentiment monitoring supports corporate social responsibility (CSR) initiatives, customer engagement, and long-term institutional reputation [8].

In Indonesia, PT Jasa Raharja as a public insurance provider faces multifaceted service challenges that make sentiment analysis particularly relevant. While its motor vehicle insurance services are generally perceived as satisfactory, customers still view the compensation claims process as less effective than expected [9]. Service quality is also shaped by organizational culture and employee motivation, where higher self-efficacy and intrinsic motivation significantly improve productivity and client interactions [10]. Recent innovations, such as online claim submissions and social media communication, have improved accessibility and engagement [11], but their long-term success requires continuous evaluation to align with user expectations. Furthermore, administrative complexity in the claims process highlights the need for an integrated one-stop service system involving multiple stakeholders, including health insurance providers and government agencies [12]. These persistent service-related issues underline the importance of leveraging social media sentiment analysis to capture public perceptions, identify pain points, and inform service innovation strategies at Jasa Raharja.

Beyond these service-related challenges, public insurance discourse in Indonesia also presents distinctive linguistic difficulties. Discussions related to insurance services are characterized not only by informal and noisy language, but also by domain-specific terminology related to claims, compensation, and service quality. These expressions often involve mixed formal–informal language, abbreviations, and context-dependent meanings. Such linguistically complex characteristics pose challenges for conventional machine learning and standard deep learning models, which may struggle to capture nuanced semantic relationships in heterogeneous social media text. Therefore, the research gap is not only methodological but also domain-specific, particularly in modelling the interaction between informal language and specialized insurance terminology. To address this challenge, this study proposes a hybrid BERT–BiLSTM–Attention model designed to capture contextual, sequential, and token-level information in Indonesian social media text within the public insurance domain.

Traditional sentiment analysis methods have relied on machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes [4], [13], supported by text mining techniques for preprocessing and feature extraction [14], [15]. While these approaches remain effective for structured domains, they often struggle with the informal, noisy, and context-dependent nature of social media discourse. For instance, although SVM demonstrates strong performance on well-structured datasets, its accuracy significantly drops when handling informal Indonesian social media texts characterized by slang, abbreviations, and code-mixing. Similarly, Naïve Bayes tends to oversimplify sentiment polarity and fails to capture contextual nuances, highlighting the need for more advanced models such as BERT that can represent semantic dependencies more effectively.

Recent advances in deep learning, particularly contextual models such as BERT, have significantly improved sentiment classification by capturing subtle linguistic cues, including irony, idioms, and semantic dependencies [16], [17], [18]. Earlier CNN and LSTM approaches improved over traditional machine learning by modeling sequential and spatial text patterns [19], [20]; nevertheless, they remain limited in capturing complex contextual semantics. In contrast, transformer-based architectures, especially BERT, achieve superior accuracy, richer representations, and greater adaptability through attention mechanisms and

large-scale pre-training [21], [22]. Pre-trained BERT models enable efficient fine-tuning with minimal labeled data, while attention further enhances interpretability by emphasizing sentiment-relevant tokens [23]. However, challenges persist in handling informal, noisy, and multilingual social media discourse [24]. Collectively, these advances have expanded applications across e-commerce, healthcare, and governance [25], [26]. However, existing transformer-based models may struggle to capture fine-grained sentiment transitions in informal Indonesian text, particularly under code-mixed and noisy conditions. In addition, standard attention mechanisms tend to distribute weights broadly, which may reduce sensitivity to critical sentiment cues. These limitations indicate the need for a systematic investigation of how contextual, sequential, and attention-based components interact in such settings. These issues highlight a methodological gap in understanding how global contextual modeling (BERT), sequential dependency modeling (BiLSTM), and attention mechanisms interact to capture sentiment transitions in informal and code-mixed Indonesian text.

Among hybrid approaches, combining BERT with Bidirectional Long Short-Term Memory (BiLSTM) networks and attention mechanisms have shown promising results, but their behavior under noisy and code-mixed Indonesian text remains insufficiently understood. BiLSTMs capture dependencies in both directions, improving contextual understanding, gradient flow, and long-term sequence [27], [28], [29]. Their versatility across domains reinforces their reliability [30], while BERT embeddings [18], [31] combined with BiLSTM sequential modelling [32] and attention for highlighting sentiment-relevant tokens [33], [34] have shown promising performance but remain insufficiently analyzed under noisy and code-mixed conditions.

Nevertheless, research applying deep learning specifically to public insurance discourse in Indonesia remains scarce. Prior studies have largely examined e-commerce or generic datasets, overlooking sector-specific challenges such as informal linguistic expressions, cultural nuances, and policy-sensitive sentiment. Moreover, existing works rarely address the policy relevance of sentiment analysis in public insurance, where public trust directly influences service adoption, compliance, and long-term sustainability. Recent studies on hybrid sentiment models in Bahasa Indonesia—such as IndoBERT combined with CNN or BiLSTM variants—have shown promising improvements, but their robustness and scalability under noisy, imbalanced, and code-mixed data remain underexplored. While previous hybrid sentiment studies focused mainly on e-commerce or product reviews, no research has yet examined the unique linguistic and policy context of Indonesia's public insurance sector. This leaves an urgent need for robust, context-aware models tailored to institutional trust monitoring and policy communication in the Indonesian public sector. Accordingly, this study addresses a key research question: How can a hybrid BERT–BiLSTM–Attention model be developed to achieve higher accuracy, robustness, and contextual relevance in analyzing public sentiment toward PT Jasa Raharja, thereby supporting institutional trust monitoring and policy communication in Indonesia's public insurance sector?

To the best of our knowledge, limited studies have explored the application of hybrid BERT–BiLSTM–Attention architectures for sentiment analysis within the context of Indonesia's public insurance domain, particularly under conditions characterized by informal, noisy, and code-mixed social media text. While prior works have demonstrated the effectiveness of combining transformer-based models with sequential and attention mechanisms, these studies largely emphasize architectural modifications or performance improvements in standard benchmark settings. In contrast, this study focuses on the systematic integration and empirical evaluation of contextual embeddings (IndoBERT), sequential modeling (BiLSTM), and attention mechanisms within a real-world, domain-specific setting. The primary contribution does not lie in proposing a novel architectural component, but in examining how these complementary mechanisms interact to improve robustness and generalization under challenging linguistic conditions, including slang, emoji usage, and class imbalance. Furthermore, the study provides a comprehensive evaluation framework that includes robustness testing under multiple perturbation scenarios, ablation analysis to quantify the contribution of each component, and statistical validation using confidence intervals and significance testing. This approach offers new empirical insights into the behavior and effectiveness of hybrid deep learning models for sentiment analysis in Indonesian social media, particularly in applications related to public service monitoring. Unlike prior works that focus on architectural innovation, this study emphasizes the systematic integration and empirical evaluation of hybrid models under real-world conditions. The main contributions of this study can be summarized as follows: (1) integration of a hybrid BERT–BiLSTM–Attention architecture for sentiment analysis in noisy and informal Indonesian social media text; (2) comprehensive and reproducible evaluation across classical machine learning, deep learning, and transformer-based baselines, supported by mean ± standard deviation, 95% confidence intervals, statistical significance testing, and ablation analysis; (3) robustness assessment under real-world conditions, including slang usage, emoji-

rich content, code-mixed Indonesian–English text, and class imbalance; and (4) application of the proposed framework for public sentiment monitoring in Indonesia's social insurance domain, providing insights for data-driven decision-making.

2. RESEARCH METHODS

This study employed a supervised deep learning approach to classify public sentiment related to Indonesia's national social insurance provider, Jasa Raharja, using a soft voting ensemble of BERT-BiLSTM-Attention models. The methodology consisted of several stages as shown in Fig. 1. The stages are data collection, preprocessing, model development and evaluation.

2.1 Data Collection

Tweets containing keywords and hashtags related to Jasa Raharja — such as "jasa raharja," "asuransi," and "kecelakaan," as well as hashtags including #JasaRaharja, #BUMNUntukIndonesia, and #SobatBUMN — were collected from Platform X (formerly Twitter) using the official Twitter API. The data collection period spans one year, from October 2022 to October 2023, allowing the dataset to capture temporal variation in public sentiment and discussion trends. A total of 4,708 public tweets written in Bahasa Indonesia were obtained. To ensure data relevance and quality, only original tweets were retained, while retweets, duplicated posts, and spam content were removed through filtering procedures. This step reduces redundancy and minimizes noise commonly found in social media data.

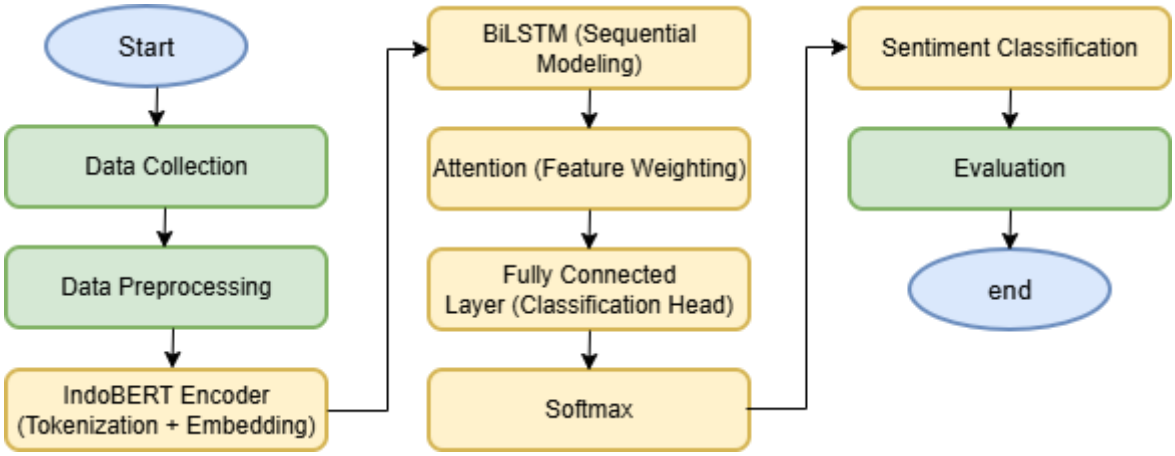


Figure 1. Research Methodology of This Research

Each record consists of three primary attributes: (i) Created at, indicating the timestamp of the post; (ii) User Name, referring to the account identifier; and (iii) Full Text, containing the tweet content. These attributes serve as the foundation for subsequent preprocessing and sentiment analysis. To support reproducibility, the data collection process is described explicitly, including keyword selection, time range, and filtering criteria. Due to platform policies and privacy considerations, the raw dataset cannot be publicly distributed. However, tweet identifiers (IDs) and the data collection protocol can be provided upon request in compliance with platform guidelines.

2.2 Data Annotation

The sentiment labeling process was conducted manually by a linguistically trained annotator with prior experience in Indonesian sentiment analysis. Before annotation, the annotator was trained using predefined guidelines and pilot examples to ensure consistent interpretation of sentiment categories. A structured annotation protocol was established, defining explicit criteria for positive, negative, and neutral labels, including guidelines for handling informal expressions, slang, sarcasm, and code-mixed Indonesian–English text. Ambiguous cases were resolved using predefined decision rules. Since a single annotator was employed, inter-annotator agreement metrics (e.g., Cohen's Kappa) could not be computed. Since a single annotator was used, no inter-annotator disagreement occurred. Instead, consistency was ensured through predefined decision rules and intra-annotator validation. To mitigate this, an intra-annotator validation procedure was performed by re-annotating a subset of the data at a different time, with results showing consistent labeling

based on the defined criteria. The dataset comprises 4,708 tweets (30.2% positive, 22.0% negative, 47.8% neutral), each including timestamp, username, text, and sentiment label (Table 1). We acknowledge that the use of a single annotator is a limitation in terms of reliability. However, the use of an expert annotator is essential for interpreting informal and code-mixed social media text. Future work will involve multiple annotators to enable inter-annotator agreement analysis and further strengthen annotation validity.

2.3 Data Preprocessing

Preprocessing is a critical step in preparing Twitter text data for sentiment analysis. Tweets are often informal and noisy, containing mentions, hashtags, URLs, emojis, and slang words. To improve the quality of the input, several steps were applied: case folding, noise removal (usernames, hashtags, URLs, numbers, and unnecessary symbols), tokenization, stopword removal, stemming (only for TIF-IDF and Word2Vec), and normalization of slang. Emojis were either removed or mapped to sentiment labels. These steps ensure that the text is clean, standardized, and compatible with the IndoBERT tokenizer, allowing the model to better capture sentiment-relevant patterns from the data.

2.3.1 Case Folding

Case folding was applied to convert all text to lowercase to reduce lexical variability in informal social media data. Although IndoBERT is case-sensitive, inconsistent capitalization in user-generated content (e.g., “Bagus”, “BAGUS”, “baGus”) may introduce unnecessary variability and fragmented representations. Standardizing text (e.g., “Bagus SEKALI!” → “bagus sekali!”) improves consistency prior to tokenization. We acknowledge that this step may remove stylistic cues such as emphasis. However, in noisy Indonesian social media text, reducing lexical variability is considered more beneficial than preserving such signals. This design reflects a trade-off between normalization and stylistic information. While no dedicated ablation was conducted for case folding, the overall robustness results indicate that this preprocessing step does not negatively impact model performance.

Table 1. Example of Dataset

Created at	User Name	Full Text	Label
Sat Oct 28 18:46:25 +0000 2023	bywellsoon	@dsyock ikut ngeluh juga teh. tanggal 19 september kemarin ayah ku meninggal dunia karena kecelakaan ,mengakibatkan pendarahan hebat dibagian kepala dan koma selama 8 hari.biaya rumah sakit ditanggung jasa raharja,kini ayah ketemu sm ibu yang udah duluan meninggal dunia	POSITIVE
Sat Oct 28 12:59:20 +0000 2023	Konsumen Cerdik	@pt_jasaraharja Bgmn caranya agar kita bebas membayar tunaidi RS jika terjadi kecelakaan kendaraan ?	NEUTRAL
Fri Oct 20 13:16:55 +0000 2023	Mdh_Sna	@pt_jasaraharja Video tertanggal 19 Oktober 2023 sekitar pukul 23:30 malam Korban langsung dibawa oleh pihak keluarga kerumah sakit Imanuel bandung Tanpa ada biaya tanpa ada arahan dari pihak yang seharusnya hadir ditengah masyarakat karena ketidaktahuan Merujuk UU Nomor 34 tahun 1964 https://t.co/Rk2T5HQ0gF	NEGATIVE

2.3.2 Noise Removal

Noise removal is applied to eliminate irrelevant elements while preserving sentiment-bearing information. Mentions (@username) and URLs are removed as they do not contribute to sentiment polarity. For hashtags, rather than removing them entirely, only the “#” symbol is stripped while retaining the associated word (e.g., “#kecewa” becomes “kecewa”). This ensures that sentiment information embedded in hashtags—such as emotional affirmation, intensification, or sentiment summarization—is preserved. Numbers and excessive symbols are selectively removed when they do not convey meaningful sentiment (e.g., reducing repetitive punctuation such as “!!!”). Overall, this step balances noise reduction with the preservation of sentiment-relevant signals, ensuring informative input for IndoBERT-based processing.

2.3.3 Tokenization

Tokenization divides text into smaller units (tokens) to enable effective language processing. In this study, tokenization was performed using the WordPiece tokenizer from the HuggingFace Transformers library to ensure compatibility with IndoBERT. For example, "saya suka produk ini" is tokenized into [saya, suka, produk, ini]. To ensure uniform input representation, a fixed maximum sequence length was applied. Sequences shorter than this length were padded with [PAD] tokens, while longer sequences were truncated using a post-truncation strategy. Padding was also applied post-sequence to maintain alignment across batches. This approach ensures consistent input dimensions and efficient model training.

2.3.4 Stopword Removal

Stopword removal is applied to eliminate frequently occurring words with limited standalone semantic value (e.g., "yang", "di", "ke", "dan") for non-transformer models such as TF-IDF, Word2Vec, and fastText. This step reduces noise and sparsity, allowing these models to focus on sentiment-bearing terms (e.g., "saya suka dengan produk ini" → "suka produk"). For transformer-based models (e.g., IndoBERT), stopword removal is not applied, as these models are pretrained on natural text where stopwords contribute to contextual and syntactic understanding. Applying stopword removal to transformers may disrupt their learned representations. To ensure a fair comparison, each preprocessing pipeline is aligned with the characteristics of the corresponding model. Specifically, traditional models rely on explicit feature representations that benefit from stopword removal, while transformer models inherently learn contextual importance, including the role of stopwords. This design reflects model-specific optimization rather than inconsistency in the experimental setup.

2.3.5 Stemming / Lemmatization

Stemming or lemmatization is applied to reduce words to their base form to address the morphological richness of the Indonesian language (e.g., "membayar" and "dibayarkan" → "bayar"). This process is performed using the Sastrawi library to improve lexical consistency and reduce vocabulary sparsity, which is particularly beneficial for non-transformer models such as TF–IDF and Word2Vec. For transformer-based models, stemming is not applied. Instead, the original text is preserved and processed using subword tokenization (WordPiece), which inherently captures morphological variations. This model-specific preprocessing ensures that each approach operates under appropriate conditions without introducing redundant transformations.

2.3.6 Slang Normalization

Slang normalization is applied to convert informal or abbreviated words commonly used in Indonesian social media into their standard forms (e.g., "gmn" → "bagaimana", "bgt" → "banget"), reducing lexical variability and improving alignment with IndoBERT's vocabulary. The normalization is performed using a predefined dictionary compiled from publicly available Indonesian slang resources and common social media usage patterns. We acknowledge that dictionary-based normalization may introduce bias, as it relies on fixed mappings and may not fully capture contextual nuances or evolving slang. To mitigate this, normalization is applied selectively to well-defined and frequently used abbreviations, while ambiguous expressions are preserved in their original form. This approach balances noise reduction with the preservation of linguistic variability.

2.3.7 Emoji and Emoticon Handling

Emoji and emoticon handling is applied to capture sentiment signals embedded in non-textual symbols commonly used in social media. In this study, emojis are not directly mapped to sentiment labels; instead, they are converted into predefined textual tokens (e.g., 😳 → "positive_emoji", 😖 → "negative_emoji") based on a standardized emoji lexicon. This approach preserves emotional information while avoiding direct label assignment that could introduce label leakage. Emojis that are ambiguous or context-dependent are retained or removed without forced mapping to prevent misrepresentation. Overall, this step enhances sentiment representation while maintaining the integrity of the learning process for IndoBERT.

2.3.8 Final Preparation

Final preparation involves transforming the cleaned text into a format compatible with IndoBERT. After preprocessing, the text is tokenized using the IndoBERT WordPiece tokenizer, which converts words or subwords into numerical token IDs corresponding to the model's vocabulary. For example, the sentence "produk bagus sekali" may be transformed into a sequence such as [101, 5678, 2345, 2021, 102]. To support transformer-based processing, an attention mask is also generated alongside the token IDs. The attention mask distinguishes valid tokens from padding tokens by assigning a value of 1 to actual tokens and 0 to padding positions. This ensures that the model focuses only on meaningful input tokens during training and ignores padded elements. This preprocessing pipeline ensures consistent input representation, enabling IndoBERT to effectively capture sentiment-related patterns and produce robust classification performance.

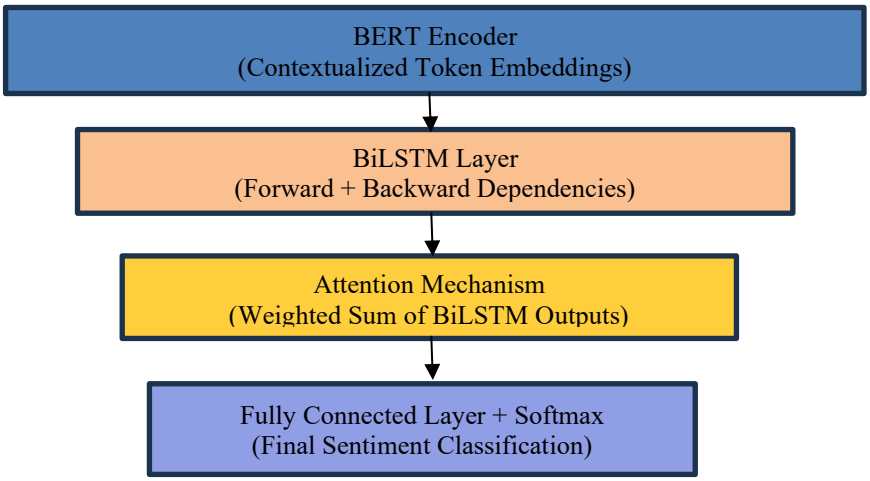


Figure 2. Hybrid BERT–BiLSTM–Attention Model

2.4. Model development

The proposed model employs a hybrid BERT–BiLSTM–Attention architecture for sentiment analysis of Indonesian social media text. While this combination is not novel, the contribution of this study lies in its systematic integration and empirical evaluation under challenging real-world conditions, including informal language, slang, code-mixing, and class imbalance. In this framework, BERT generates contextual embeddings, which are further processed by a BiLSTM layer to capture sequential dependencies and sentiment transitions not fully modelled by transformer representations alone. An additive attention mechanism is then applied to emphasize sentiment-relevant tokens, improving focus on informative parts of the sequence. To enhance robustness and generalization, multiple configurations are evaluated using varying hyperparameters rather than relying on a single setting. Unlike prior work that emphasizes architectural novelty, this study focuses on robustness-oriented design and empirical validation. The contributions are threefold: (1) analysing the complementary roles of sequential modelling and attention in transformer-based representations, (2) evaluating model robustness under realistic perturbations, and (3) demonstrating stable performance across multiple configurations. The overall framework is illustrated in Fig. 2.

2.4.1. BERT Encoder

The BERT encoder is the first stage of the proposed hybrid model. In this study, the IndoBERT-base model is employed as the contextual feature extractor. After preprocessing and tokenization, each tweet is split into WordPiece sub-tokens and mapped to vocabulary IDs. For each position, BERT constructs an input representation by summing token, segment, and positional embeddings, which are then processed through a stack of Transformer encoder layers. Within each layer, multi-head self-attention enables each token to attend to all others, producing contextualized embeddings that capture semantic and syntactic dependencies across the sequence. Trained on large-scale Indonesian corpora, IndoBERT-base provides rich representations suitable for downstream sentiment analysis. In this architecture, BERT serves as a feature extractor, where the token-level sequence outputs from the final encoder layer (as defined in Eq. (1)) are used as inputs to the subsequent BiLSTM layer for sequential modelling.

$$H^{(L)} = [h_1; \ldots; h_n] \in \mathbb{R}^{n \times d}, \quad (1)$$

where n is the number of tokens. d is the embedding dimension. Each h_i already encodes contextual semantic and syntactic information.

2.4.2. BiLSTM Layer

The contextual embeddings generated by BERT are subsequently passed to a BiLSTM layer to capture sequential dependencies. Although BERT is inherently bidirectional, it models global context through parallel self-attention. In contrast, BiLSTM explicitly captures sequential and local context transitions, which are particularly relevant in informal Indonesian text, such as negation patterns and multi-clause expressions. To validate its contribution, an ablation study (Table 6) is conducted, where removing the BiLSTM layer results in a performance decrease. This indicates that BiLSTM provides complementary information rather than redundant processing.

Forward LSTM

It processes the embeddings from the first token to the last token, from h_1 to h_n, capturing past-to-future dependencies as defined in Eq. (2).

$$h_t^{fwd} = LSTM(h_{t-1}, x_t). \quad (2)$$

In Eq. (2), h_t denotes the hidden state at time step t, representing the updated contextual information after processing the current token. The term h_{t-1} is the hidden state from the previous time step, which carries information from all preceding tokens and enables the model to capture past dependencies. The variable x_t refers to the input embedding of the t-th token, containing its semantic and syntactic features. The function $LSTM(\cdot)$ takes the previous hidden state and the current input embedding to compute the new hidden state through its internal gating mechanisms.

Backward LSTM

It processes the embeddings in reverse, from h_n to h_1, capturing future-to-past dependencies as defined in Eq. (3).

$$h_t^{bwd} = LSTM(h_{t+1}, x_t). \quad (3)$$

Concatenation

At each time step t, the forward and backward hidden states are concatenated to form a unified and context-enriched token representation, as expressed in Eq. (4).

$$h_t^{BiLSTM} = [h_t^{fwd}; h_t^{bwd}] \in \mathbb{R}^{2d_{lstm}}. \quad (4)$$

As a result, the BiLSTM layer produces a new sequence of representations, as defined in Eq. (5).

$$H^{BiLSTM} = [h_1^{BiLSTM}, h_2^{BiLSTM}, \ldots, h_n^{BiLSTM}]. \quad (5)$$

This combination ensures that the model does not only rely on BERT's context-aware embeddings but also incorporates explicit sequential order information, which is especially important for languages like Bahasa Indonesia where word order and negations (e.g., "tidak bagus" vs "bagus") can drastically change sentiment meaning. Thus, BiLSTM enriches the BERT embeddings with directional temporal context, preparing them for the subsequent attention mechanism to highlight the most sentiment-relevant words.

2.4.3. Attention Mechanism

On top of BiLSTM outputs, an attention mechanism is after passing through the BiLSTM, each token in the sequence is represented by a hidden state that combines both forward and backward contextual information, as represented in Eq. (6).

$$H^{BiLSTM} = [h_1^{BiLSTM}, h_2^{BiLSTM}, \ldots, h_n^{BiLSTM}]. \quad (6)$$

While these representations are rich, not every word in a tweet contributes equally to determining sentiment. For example, in the sentence *"Pelayanan cepat tetapi mahal"*, the word *"mahal"* carries more weight in shaping the negative sentiment than *"cepat"*. To address this, we apply an **attention mechanism**, which dynamically assigns importance scores to each token.

The attention process works as follows:

1. Intermediate Representation

   Each BiLSTM output is projected into a hidden space using a nonlinear transformation, as defined in Eq. (7).

$$u_i = tanh(W_a h_i^{BiLSTM} + b_a), \tag{7}$$

   where W_a and b_a are trainable parameters, and u_i is the intermediate representation of token i.

2. Attention Weights

   A similarity score is computed between each u_i and a learnable context vector w_a. This score is then normalized across all tokens using a Softmax function which is defined by Eq. (8).

$$\alpha_i = \frac{exp(u_i^T W_a)}{\sum_{j=1}^{n} exp(u_i^T W_a)}. \tag{8}$$

   The resulting α_i represents the attention weight for token i, i.e., its relative importance in the sequence.

3. Weighted Sentence Representation

   The final sentence vector s is obtained as a weighted sum of the BiLSTM hidden states, as defined in Eq. (9).

$$s = \sum_{i=1}^{n} \alpha_i h_i^{BiLSTM}. \tag{9}$$

   This mechanism ensures that sentiment-relevant words receive higher weights while less informative words contribute less to the final representation. As a result, the model can focus more effectively on the parts of the tweet that determine its overall sentiment.

The output vector s is then passed into a fully connected classification layer with a softmax function to produce the final sentiment label (positive, negative, or neutral). The Softmax function is an activation function that converts a vector of real numbers (called logits) into a probability distribution, where all probabilities add up to 1. For an input vector $z = [z_1, z_2, \ldots, z_K]$, the Softmax function is defined as Eq. (10).

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}}, \tag{10}$$

where e^{z_i} is the exponential of the i-th score. The denominator sums all exponentials, ensuring probabilities sum to 1. $\sigma(z_i)$: the predicted probability for class i.

Although the attention mechanism used in this study follows a standard additive formulation, it is employed to provide token-level weighting over BiLSTM outputs, enabling the model to emphasize sentiment-relevant words. This is particularly important in informal Indonesian social media text, where sentiment is often concentrated in specific tokens. An ablation study (Table 6) shows that removing the attention layer leads to a decrease in performance, indicating its complementary contribution rather than redundancy.

2.5. Model Evaluation

The evaluation framework was designed to assess performance, robustness, and generalization of the proposed Hybrid BERT–BiLSTM–Attention model. Four metrics were used: Accuracy, Macro Precision, Macro Recall, and Macro F1-score to ensure balanced evaluation under class imbalance. Results are reported as mean ± standard deviation over five runs, with 95% confidence intervals to capture variability. ROC curves

and AUC values were computed per class, while statistical significance was evaluated using paired t-tests ($p < 0.05$). Additionally, effect size (Cohen's d) is reported to quantify the magnitude of performance differences. Further analysis includes ablation studies to assess the contribution of BERT, BiLSTM, and attention components, learning curves to evaluate sample efficiency, and robustness testing under controlled perturbations (e.g., slang, elongation, emoji, code-mixing, and class imbalance). Performance degradation was measured relative to the clean baseline. To ensure reproducibility, the dataset was split into training and testing sets (80:20) with a fixed random seed. Experiments were repeated five times with different seeds, and the test set was strictly held out to prevent data leakage. All evaluation procedures were applied consistently across models to ensure fair comparison.

Table 2. Hyperparameters Setup of BERT, BiLSTM and Attention

Component	Hyperparameter	Hyperparameter values of This Research	Typical Range / Value based on References	Notes / Recommended Choice	References
BERT	Learning Rate	2e-5	1e-5 – 5e-5	Optimal values often around 2e-5 or 3e-5	[35], [36]
	Batch Size	32 (train)	16 – 128	32 or 64 commonly used; larger sizes if GPU memory allows	[37], [38]
	Epochs	5	3 – 10	4–5 sufficient to prevent overfitting	[36]
	Max Sequence Length	256	128 – 512	256 for efficiency, 512 for full context	[37]
	Dropout Rate	0.1	0.1 – 0.3	0.1 is default in BERT-Base	[39]
	Warm-up Steps	10% of total steps	10–20% of total steps	Helps stabilize training at the start	[40]
	Weight Decay	0.01	0.01 – 0.1	Commonly set at 0.01	[41]
	Optimizer	AdamW	Adam / AdamW	AdamW preferred for better generalization	[42]
BiLSTM	Learning Rate	1e-4	1e-3 – 1e-5	Typically 1e-4 for sentiment and time series tasks	[43], [44]
	Batch Size	32	16 – 64	32 is widely adopted	[44]
	Epochs	25	10 – 50	20–30 epochs usually optimal; monitor validation loss	[45], [46]
	Max Sequence Length	128 (token-level for RNN input)	50 – 200 tokens	Adjusted to dataset, often 100–128 in NLP tasks	[47]
	Dropout Rate	0.3	0.2 – 0.5	0.3 frequently reported as effective	[48]
Attention	Learning Rate	1e-4 (same optimizer as BiLSTM)	1e-5 – 1e-3	1e-4 or 5e-5 yields stable training	[49], [50]
	Batch Size	32	16 – 64 (up to 128/256)	32–64 optimal; larger sizes if resources permit	[51]
	Epochs	25	10 – 50	20–40 to capture complex relationships	[52]
	Dropout Rate	0.3	0.1 – 0.5	0.3 balances generalization and performance	[50], [52]
	Attention Heads	N/A (additive attention)	4 – 16	8 commonly adopted	[50], [53]
	Model Dimension (d_model)	256	128 – 512	256 often used as a starting point	[52], [54]
	Sequence Length	128	50 – 512 tokens	128–256 for efficiency, 512 for full coverage	[49]
	Weight Decay	1e-4	1e-5 – 1e-2	1e-4 often recommended	[51]
	Optimizer	Adam	Adam / AdamW	Adam widely used for attention-based models	[55]

2.6. Experimental Setup

The experimental setup was designed to evaluate the proposed hybrid BERT–BiLSTM–Attention model in a controlled and reproducible environment. The workflow consists of four stages: data collection, preprocessing, model construction, and evaluation. A total of 4,708 Indonesian tweets related to PT Jasa Raharja were collected via the Twitter API, retaining only original posts after removing retweets and spam. Each record includes timestamp, username, text, and sentiment label (positive, negative, neutral). Preprocessing involved case folding, noise removal (mentions, URLs, symbols), slang normalization, and emoji handling. For non-transformer models, additional steps such as tokenization, stopword removal, and stemming (Sastrawi) were applied. In contrast, transformer-based models used the IndoBERT WordPiece tokenizer without stopword removal or stemming to preserve contextual information. The proposed model integrates IndoBERT embeddings with BiLSTM and an attention mechanism. The dataset was split into training and testing sets (80:20) using a fixed random seed (42). Multiple configurations were evaluated under consistent hyperparameter settings (Table 2) to ensure fair comparison. Performance metrics were computed on the held-out test set. Three model families were benchmarked: (i) traditional machine learning (LogReg, SVM, NB with TF–IDF), (ii) deep learning (LSTM, BiLSTM, TextCNN with Word2Vec/fastText), and (iii) transformer-based models (IndoBERT and mBERT). All experiments were implemented in Python (PyTorch, Transformers, PyTorch Lightning) and executed on an NVIDIA RTX 3060 GPU with mixed precision (FP16) to ensure computational efficiency and reproducibility.

3. RESULTS AND DISCUSSIONS

In this section, the performance of the proposed Hybrid BERT–BiLSTM–Attention model is evaluated and compared against multiple baseline approaches, including traditional machine learning classifiers, deep learning architectures without transformers, and transformer-based models. As summarized in Table 3, the proposed hybrid model consistently outperforms all baselines in terms of classification accuracy, robustness to noisy data, and statistical reliability, confirming its effectiveness for sentiment analysis in the context of Indonesia's public insurance sector.

Table 3. Performance Comparison of Models on the Test Set
(Mean ± Standard Deviation and 95% Confidence Interval over Five Runs)

Model	Accuracy	95% CI	Macro Precision	95% CI	Macro Recall	95% CI	Macro F1 Score	95% CI
Logistic Regression (TF–IDF)	0.72 ± 0.015	0.701–0.739	0.71 ± 0.017	0.689–0.731	0.71 ± 0.016	0.690–0.730	0.71 ± 0.016	0.690–0.730
Linear SVM (TF–IDF)	0.74 ± 0.013	0.724–0.756	0.74 ± 0.014	0.723–0.757	0.73 ± 0.015	0.711–0.749	0.74 ± 0.014	0.723–0.757
Multinomial Naïve Bayes (TF–IDF)	0.70 ± 0.018	0.678–0.722	0.69 ± 0.019	0.666–0.714	0.70 ± 0.018	0.678–0.722	0.69 ± 0.019	0.666–0.714
LSTM (Word2Vec)	0.77 ± 0.012	0.755–0.785	0.77 ± 0.013	0.754–0.786	0.76 ± 0.012	0.745–0.775	0.77 ± 0.012	0.755–0.785
BiLSTM (Word2Vec)	0.78 ± 0.011	0.766–0.794	0.78 ± 0.012	0.765–0.795	0.77 ± 0.013	0.754–0.786	0.78 ± 0.012	0.765–0.795
TextCNN (fastText)	0.77 ± 0.012	0.755–0.785	0.77 ± 0.013	0.754–0.786	0.76 ± 0.014	0.743–0.777	0.77 ± 0.013	0.754–0.786
IndoBERT (fine-tuned)	0.83 ± 0.010	0.818–0.842	0.83 ± 0.011	0.816–0.844	0.82 ± 0.012	0.805–0.835	0.83 ± 0.011	0.816–0.844
Hybrid BERT–BiLSTM–Attention	**0.85 ± 0.009**	**0.839–0.861**	**0.85 ± 0.010**	**0.838–0.862**	**0.84 ± 0.011**	**0.826–0.854**	**0.85 ± 0.010**	**0.838–0.862**

Note: Results are reported as mean ± standard deviation over five independent runs. The 95% confidence intervals are computed using the t-distribution (n = 5, t = 2.776).

The proposed hybrid BERT–BiLSTM–Attention model was benchmarked against three categories of baselines: (i) traditional machine learning models with TF–IDF features, (ii) deep learning models using Word2Vec and fastText embeddings, and (iii) transformer-based IndoBERT. As shown in Table 3, performance improves progressively across model families. Traditional models achieve accuracy between 0.70–0.74 (e.g., Logistic Regression: 0.72, 95% CI: 0.701–0.739), reflecting limited contextual

understanding. Deep learning models improve performance to 0.77–0.78 (e.g., BiLSTM: 0.78, 95% CI: 0.766–0.794) by capturing sequential patterns, but remain constrained by static embeddings. IndoBERT further enhances performance (Accuracy = 0.83, 95% CI: 0.818–0.842) through contextual representations. The hybrid model achieves the highest performance (Accuracy = 0.85, Macro-F1 = 0.85), demonstrating the benefit of integrating contextual (BERT), sequential (BiLSTM), and attention-based representations. Tables 3 and 4 confirm consistent improvements across all metrics. These gains are supported by confidence intervals and statistical tests, showing significant improvements over traditional and deep learning models ($p < 0.001$) and a reliable improvement over IndoBERT ($p = 0.027$). This performance advantage reflects the complementary roles of the model components: BERT captures contextual semantics, BiLSTM models sequential dependencies, and attention emphasizes sentiment-relevant tokens. Together, they enable more robust and precise sentiment classification in informal and context-dependent Indonesian social media.

Table 4. Statistical Significance of Macro-F1 Score Differences (Test Set).

Comparison	Macro F1 (A)	Macro F1 (B)	p-value
Hybrid vs LogReg	0.85	0.71	<0.001
Hybrid vs SVM	0.85	0.74	<0.001
Hybrid vs Naïve Bayes	0.85	0.69	<0.001
Hybrid vs LSTM	0.85	0.77	0.003
Hybrid vs BiLSTM	0.85	0.78	0.004
Hybrid vs TextCNN	0.85	0.77	0.004
Hybrid vs IndoBERT	0.85	0.83	0.027

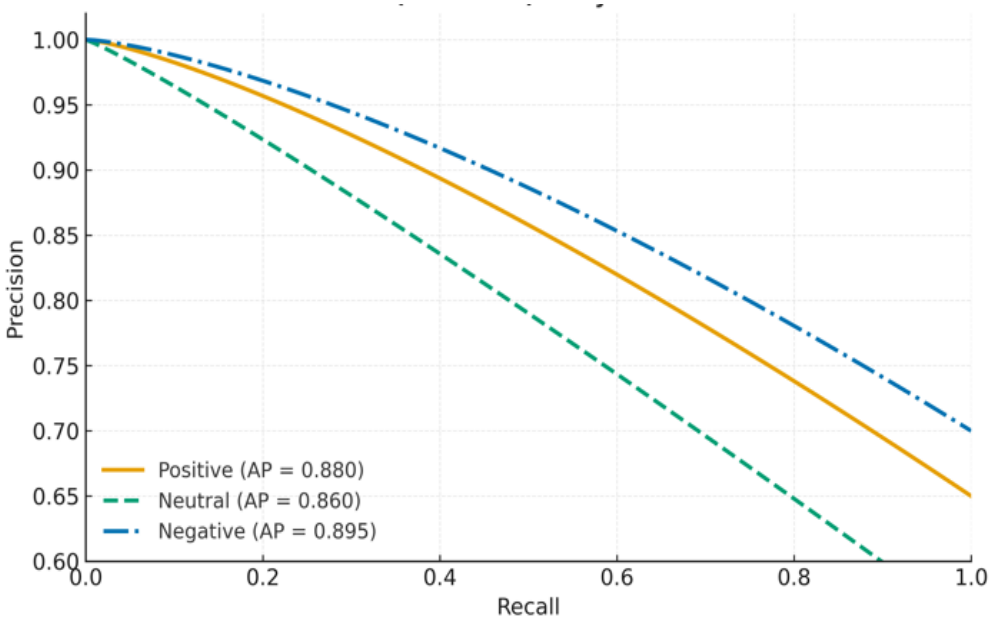


Figure 3. ROC Curves (One-vs-Rest) for Each Sentiment Class Resulted by a Hybrid BERT-BILSTM-Attention Model

The Hybrid BERT–BiLSTM–Attention model demonstrates strong discriminative performance across sentiment classes, as shown in Fig. 3. The Precision–Recall curves yield high Average Precision (AP) scores: 0.880 (positive), 0.860 (neutral), and 0.895 (negative), indicating consistent precision across varying recall levels, even under class imbalance and noisy conditions. The negative class achieves the highest AP, reflecting the model's effectiveness in capturing explicit complaint-related expressions, while the lower performance for the neutral class highlights its inherent ambiguity and context dependence. This performance is driven by the complementary roles of the model components: BERT captures contextual semantics, BiLSTM models sequential dependencies, and attention emphasizes sentiment-relevant tokens. Together, these enable the model to effectively capture both explicit and implicit sentiment cues, including negation and multi-clause structures. Overall, the results confirm that the hybrid architecture provides robust and balanced performance, making it well-suited for sentiment analysis in noisy and informal Indonesian social media.

Table 5. Training Cost Summary (Mean±SD over 3 Runs)

Model	#Params (Million)	Train / epoch (min)	Total train (min)
Logistic Regression (TF–IDF)	N/A	0.6 ± 0.1	3.0 ± 0.2
Linear SVM (TF–IDF)	N/A	0.9 ± 0.1	4.5 ± 0.3
Multinomial Naïve Bayes	N/A	0.3 ± 0.1	1.5 ± 0.1
BiLSTM (Word2Vec)	2.8	3.1 ± 0.2	18.5 ± 0.6
TextCNN (fastText)	2.1	2.4 ± 0.2	14.2 ± 0.5
IndoBERT (fine-tuned)	109	7.5 ± 0.4	44.0 ± 1.2
Hybrid BERT–BiLSTM–Attention	111	9.0 ± 0.5	52.0 ± 1.5

As shown in Table 5, computational cost increases progressively across model families. TF–IDF–based classifiers are the most efficient (1.5–4.5 minutes) due to their simple linear structure, while neural models such as BiLSTM and TextCNN require longer training times (14.2–18.5 minutes) due to sequential and convolutional operations. Transformer-based models, including IndoBERT and the proposed hybrid model, incur substantially higher costs given their large parameter sizes (109M–111M). The hybrid BERT–BiLSTM–Attention model requires the longest training time (52.0 ± 1.5 minutes), reflecting the added complexity of sequential and attention layers. Despite this, it achieves the highest performance (Accuracy = 0.85; Macro-F1 = 0.85) and demonstrates improved robustness and generalization. These results highlight a clear performance–efficiency trade-off: increased model complexity leads to higher computational cost but yields measurable gains in classification quality. This makes the hybrid model suitable for applications where accuracy and reliability are prioritized over computational efficiency, particularly in noisy and complex Indonesian social media settings.

Table 6. Ablation of Model Components (Test Set)

Model Variant	Accuracy	Macro-F1	Δ F1	Δ (%)	Interpretation
Full Model (BERT + BiLSTM + Attention)	0.84	0.82	—	—	Baseline model
w/o BiLSTM	0.82	0.80	−0.02	−2.44%	Loss of sequential transition modeling
w/o Attention	0.83	0.81	−0.01	−1.22%	Reduced focus on sentiment-critical tokens

Table 6 presents the ablation results to assess the contribution of each component in the proposed hybrid architecture. The full model achieves the highest performance (Macro-F1 = 0.82), serving as the baseline for comparison. Removing the BiLSTM layer results in a performance drop of 0.02 (−2.44%), indicating its important role in capturing sequential dependencies and sentiment transitions across clauses. This finding aligns with the nature of informal Indonesian social media text, where sentiment is often expressed through multi-clause structures and contrastive patterns. Similarly, excluding the attention mechanism leads to a decrease of 0.01 (−1.22%) in Macro-F1, demonstrating its contribution in emphasizing sentiment-critical tokens. Without attention, the model tends to distribute importance more uniformly across tokens, reducing its ability to focus on decisive sentiment cues. Overall, the results show that both BiLSTM and attention contribute complementary benefits. While BiLSTM enhances the modeling of sequential sentiment transitions, attention improves token-level discrimination. The performance degradation observed in both ablated variants confirms that the proposed architecture does not rely on redundant components, but instead benefits from the interaction between contextual (BERT), sequential (BiLSTM), and attention-based representations. These findings provide strong empirical evidence supporting the necessity of integrating BiLSTM and attention layers, particularly for handling complex sentiment patterns in informal and noisy Indonesian text. While this study focuses on component-level ablation, further exploration of architectural variations such as BiLSTM depth and attention configurations remains an important direction for future work.

Table 7. Robustness Evaluation under Noisy and Imbalanced Conditions

Scenario	IndoBERT (fine-tuned)	Hybrid
Baseline (clean)	0.83 ± 0.011	0.85 ± 0.010
+ Slang/abbreviation injection	0.81 ± 0.012	0.83 ± 0.011
+ Character elongation (≤3 repetitions)	0.81 ± 0.012	0.83 ± 0.011
+ Emoji and hashtag intensive	0.82 ± 0.012	0.84 ± 0.011
Code-mix (ID–EN subset)	0.80 ± 0.013	0.83 ± 0.012
Minority downsampling (negative −30%)	0.79 ± 0.013	0.82 ± 0.012

To evaluate robustness under realistic noisy conditions, several inference-time perturbations were introduced, including slang insertion, character elongation, emoji and hashtag injection, code-mixing (Indonesian–English), and class imbalance. As shown in Table 7, both IndoBERT and the Hybrid BERT–BiLSTM–Attention model exhibit only modest performance degradation (0.02–0.03 Macro-F1). Notably, the hybrid model consistently outperforms IndoBERT across all scenarios, indicating stronger robustness to linguistic variations. This improvement is driven by the complementary integration of model components. While IndoBERT captures contextual semantics, it does not explicitly model sequential dependencies. The addition of BiLSTM enables better handling of word-order and context-dependent patterns, such as negation. The attention mechanism further enhances performance by emphasizing sentiment-relevant tokens and reducing the influence of noise. Overall, these results demonstrate that the hybrid model maintains stable performance under diverse perturbations, highlighting its suitability for real-world sentiment analysis in noisy Indonesian social media.

Table 8. Out-of-Time Evaluation: Temporal Generalization across Consecutive Periods

Model	In-time F1	OOT F1	Δ (Absolute / Relative Change)
IndoBERT (fine-tuned)	0.83 ± 0.011	0.79 ± 0.013	−0.04 / −4.8%
Hybrid	0.85 ± 0.010	0.82 ± 0.012	−0.03 / −3.5%

To assess temporal generalization, an Out-of-Time (OOT) evaluation was conducted by training the models on data from month t and testing on unseen data from month $t + 1$, thereby simulating temporal drift in sentiment expression. As summarized in Table 8, both IndoBERT and the Hybrid BERT–BiLSTM–Attention model exhibit performance degradation under this setting, reflecting the impact of evolving language use, topical shifts, and changes in sentiment distribution over time. Specifically, IndoBERT's Macro-F1 decreases by 0.04 (−4.8%), while the hybrid model shows a smaller reduction of 0.03 (−3.5%), indicating improved resilience to temporal drift. This result suggests that the integration of sequential modeling and attention mechanisms enables the hybrid architecture to better adapt to shifts in linguistic patterns compared to transformer-only approaches. Overall, these findings demonstrate that the proposed hybrid model maintains more stable performance under temporally shifted data, highlighting its potential for real-world sentiment analysis scenarios where data distributions evolve over time.

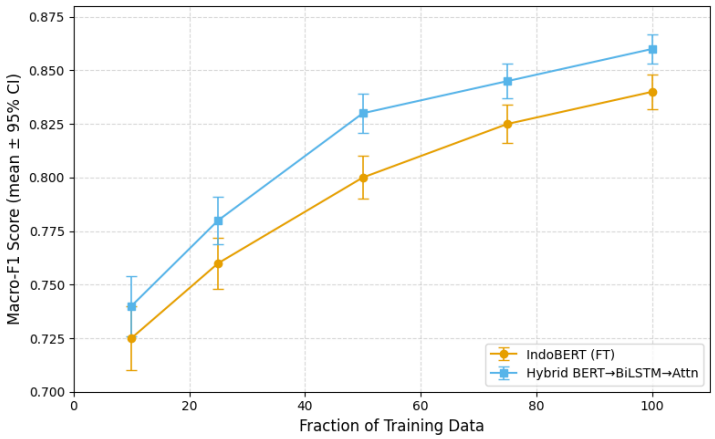


Figure 4. Learning Curves (Macro-F1 Score) vs. Fraction of Training Data.

As illustrated in Fig. 4, model performance improves consistently with increasing training data. Across all data fractions, the Hybrid BERT–BiLSTM–Attention model outperforms IndoBERT, indicating better learning and generalization under limited data conditions. Notably, the hybrid model achieves approximately 0.84 Macro-F1 using only 75% of the data, suggesting strong sample efficiency. This can be attributed to the complementary integration of contextual embeddings (BERT), sequential modeling (BiLSTM), and attention mechanisms, which enhance semantic understanding, capture dependencies, and emphasize sentiment-relevant tokens. These results highlight the model's robustness, particularly in scenarios where labeled data are limited.

As shown in Table 9, performance improves progressively from traditional machine learning to deep learning and transformer-based models. Classical methods (LogReg, SVM, NB) rely on surface features and are prone to errors in negation and contrast (e.g., "Gratis tapi antrian parah"). While BiLSTM and TextCNN improve sequential modeling, they remain limited by static embeddings. IndoBERT provides stronger

contextual understanding but may struggle with multi-clause sentiment structures. In contrast, the hybrid model better captures sentiment transitions, as seen in “Pelayanan mantap... nunggu lama”, where IndoBERT predicts positive while the hybrid model correctly identifies negative sentiment. This improvement is driven by the complementary roles of BiLSTM and attention, resulting in more robust handling of complex sentiment patterns.

Table 9. Prediction Comparison Across Model Architectures

No	Text	True Label	Log Reg	SVM	NB	Bi-LSTM	Text-CNN	Indo-BERT	Hybrid	Key Observation
1	Pelayanan cepat dan membantu	POS	POS	POS	POS	POS	POS	POS	POS	All models correctly capture clear positive sentiment
2	Bayar mahal tapi pelayanan lama	NEG	NEG	NEG	POS	NEG	NEG	NEG	NEG	NB fails due to word independence assumption
3	Kadang bagus kadang ribet	NET	NEG	NET	NEG	NET	NET	NET	NET	Traditional models struggle with mixed sentiment
4	Lumayan tapi perlu ditingkatkan	NET	POS	POS	POS	NET	POS	NET	NET	Word-based models biased toward “lumayan”
5	Gratis tapi antrian parah	NEG	POS	NEG	POS	NEG	NEG	NEG	NEG	LogReg & NB biased toward “gratis”
6	Tidak buruk tapi tidak bagus	NET	NEG	NEG	NEG	NET	NEG	NET	NET	Classical models fail on negation patterns

The proposed hybrid model introduces higher computational complexity due to the integration of transformer, sequential, and attention components, reflecting a trade-off between performance and efficiency. While the model achieves improved accuracy and robustness, inference latency was not explicitly measured, which represents a limitation. Given its architectural complexity, inference time is expected to be higher than that of simpler baselines. Future work will include detailed analysis of computational efficiency, including latency and scalability for real-time deployment. Despite its promising performance, this study has several limitations. The relatively small dataset (4,708 tweets) may limit generalizability, and the higher computational cost may pose challenges for large-scale or real-time applications. Future work will focus on expanding datasets, improving efficiency, and exploring advanced architectures, including benchmarking against recent large language models (LLMs) and integrating parameter-efficient techniques to enhance scalability and performance.

4. CONCLUSION

This study proposes a hybrid BERT–BiLSTM–Attention model for sentiment analysis of Indonesian social media in the public insurance domain. The results show consistent performance improvements over traditional, deep learning, and transformer-based baselines (Accuracy = 0.85, Macro-F1 = 0.85), with statistical validation (confidence intervals, significance testing, and effect size) confirming the reliability of these gains. The model also demonstrates robustness under noisy, code-mixed, and evolving data conditions. The contribution of this work lies not in architectural novelty, but in the systematic integration and empirical evaluation of contextual, sequential, and attention-based representations. The findings indicate that these components provide complementary benefits, particularly in capturing sentiment transitions and multi-clause structures in informal text. Practically, the model supports data-driven sentiment monitoring for public institutions. However, limitations include a relatively small dataset, higher computational cost, the absence of inference latency analysis, and a component-level (rather than architectural) ablation study. Future work will address these aspects by expanding datasets, evaluating computational efficiency, and exploring

advanced architectural configurations, including LLM-based and parameter-efficient approaches. Overall, this study provides empirical insights into the effectiveness of hybrid models for sentiment analysis in noisy Indonesian social media.

Author Contributions

Syaiful Anam: Conceptualization, Methodology, Original Draft Writing and Funding Acquisition. Nur Atiqah Sia Abdullah: Review, Editing and Validation. Hilmi Aziz Bukhori: Investigation, Software Implementation and Visualization. Avin Maulana: Formal Analysis, Resources and Project Administration. Regina Vincentia: Data Curation, Resource Provision. All authors discussed the results collaboratively and contributed to the final version of the manuscript.

Funding Statement

This research received no external funding and was partially supported by internal resources provided by the Brawijaya University.

Acknowledgment

The authors would like to express their sincere gratitude to the Visiting Lecturer (VL) Program of Universitas Brawijaya (UB) and the faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia, for their academic collaboration and support in completing this research.

Declarations

The authors confirm that there are no conflicts of interest associated with the research, authorship, or publication of this article.

Declaration of Generative AI and AI-assisted technologies

The authors used ChatGPT only to assist with language polishing and formatting consistency, including improving wording and ensuring uniform terminology. No AI was used to generate research content, perform analyses, or create or modify figures and tables. The authors have fully reviewed the manuscript and take full responsibility for its content.

REFERENCES

[1] P. W. Handayani, G. A. Zagatti, H. Kéfi, and S. Bressan, "IMPACT OF SOCIAL MEDIA USAGE ON USERS' COVID-19 PROTECTIVE BEHAVIOR: SURVEY STUDY IN INDONESIA," 2023. doi: https://doi.org/10.2196/46661.

[2] C. A. de Castro, "THEMATIC ANALYSIS IN SOCIAL MEDIA INFLUENCERS: WHO ARE THEY FOLLOWING AND WHY?," 2023. doi: https://doi.org/10.3389/fcomm.2023.1217684.

[3] S. Bin Naeem and M. N. K. Boulos, "COVID-19 MISINFORMATION ONLINE AND HEALTH LITERACY: A BRIEF OVERVIEW," 2021. doi: https://doi.org/10.3390/ijerph18158091.

[4] Z. Zakaria, K. Kusrini, and D. Ariatmanto, "SENTIMENT ANALYSIS TO MEASURE PUBLIC TRUST IN THE GOVERNMENT DUE TO THE INCREASE IN FUEL PRICES USING NAIVE BAYES AND SUPPORT VECTOR MACHINE," *International Journal of Artificial Intelligence & Robotics (Ijair)*, 2023. doi: https://doi.org/10.25139/ijair.v5i2.7167.

[5] N. Ma, G. Yu, X. Jin, and X. Zhu, "QUANTIFIED MULTIDIMENSIONAL PUBLIC SENTIMENT CHARACTERISTICS ON SOCIAL MEDIA FOR PUBLIC OPINION MANAGEMENT: EVIDENCE FROM THE COVID-19 PANDEMIC," *Front. Public Health*, 2023. doi: https://doi.org/10.3389/fpubh.2023.1097796.

[6] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, and M. A. Shah, "IMPROVING SENTIMENT ANALYSIS IN SOCIAL MEDIA BY HANDLING LENGTHENED WORDS," *Ieee Access*, 2023. doi: https://doi.org/10.1109/ACCESS.2023.3238366

[7] M. Slijepčević, N. P. Šević, G. Ašanin, and M. Ilić, "INFLUENCE OF CSR ACTIVITIES ON OPINIONS OF INSURANCE SERVICE USERS WHEN PURCHASING INSURANCE," *Tokovi Osiguranja*, 2023. doi: https://doi.org/10.5937/TokOsig2302202S.

[8] J. Park, W. Choi, and S.-U. Jung, "EXPLORING TRENDS IN ENVIRONMENTAL, SOCIAL, AND GOVERNANCE THEMES AND THEIR SENTIMENTAL VALUE OVER TIME," 2022. doi: https://doi.org/10.3389/fpsyg.2022.890435.

[9] F. Rozy, H. Harmain, and J. Nasution, "ANALYSIS OF THE EFFECTIVENESS OF MOTOR VEHICLE INSURANCE AT PT. JASA RAHARJA," 2023. doi: https://doi.org/10.62398/probis.v14i6.384.

[10] N. Phan Thi Hang and N. Kim Quoc Trung, "SERVICE QUALITY, CUSTOMER SATISFACTION AND LOYALTY: A CASE STUDY IN VIETNAMESE SMES," *Cogent Business and Management*, vol. 11, no. 1, 2024. doi: https://doi.org/10.1080/23311975.2024.2377769.

[11] S. Sinta and S. Suyanto, "KOMUNIKASI INOVASI JASA RAHARJA CABANG RIAU DALAM MENGIMPLEMENTASIKAN PELAYANAN SANTUNAN ONLINE KORBAN KECELAKAAN DI KOTA PEKANBARU," 2024. doi: https://doi.org/10.62383/aktivisme.v1i4.469.
[12] M. Janssen and G. Kuk, "THE CHALLENGES AND LIMITS OF BIG DATA ALGORITHMS IN TECHNOCRATIC GOVERNANCE," *Gov. Inf. Q.*, vol. 33, no. 3, pp. 371–377, 2016. doi: https://doi.org/10.1016/j.giq.2016.08.011.
[13] M. Rifaí, R. Buaton, and I. G. Prahmana, "SENTIMENT ANALYSIS USING TEXT MINING TECHNIQUES ON SOCIAL MEDIA USING THE SUPPORT VECTOR MACHINE METHOD CASE STUDY SEAGAMES 2023 FOOTBALL FINAL," *J. Of Artif. Intell. And Eng. Appl.*, 2023. doi: https://doi.org/10.59934/jaiea.v3i1.274.
[14] N. Z. Al Habesyah, R. Herteno, F. Indriani, I. Budiman, and D. Kartini, "SENTIMENT ANALYSIS OF TIKTOK SHOP CLOSURE IN INDONESIA ON TWITTER USING SUPERVISED MACHINE LEARNING," *Journal of Electronics Electromedical Engineering and Medical Informatics*, 2024. doi: https://doi.org/10.35882/jeeemi.v6i2.381.
[15] S. Rahman, N. Jahan, F. Sadia, and I. Mahmud, "SOCIAL CRISIS DETECTION USING TWITTER BASED TEXT MINING-A MACHINE LEARNING APPROACH," *Bulletin of Electrical Engineering and Informatics*, 2023. doi: https://doi.org/10.11591/eei.v12i2.3957.
[16] A. D. Herlia and M. Andriansyah, "SENTIMENT ANALYSIS OF FLO APPLICATIONS FOR WOMEN'S NEEDS USING THE CNN AND LSTM ALGORITHMS," *Eduvest - Journal of Universal Studies*, 2024. doi: https://doi.org/10.59188/eduvest.v4i5.1182.
[17] D. Olaniyan, R. O. Ogundokun, O. P. Bernard, J. Olaniyan, R. Maskeliūnas, and H. B. Akande, "UTILIZING AN ATTENTION-BASED LSTM MODEL FOR DETECTING SARCASM AND IRONY IN SOCIAL MEDIA," *Computers*, 2023. doi: https://doi.org/10.3390/computers12110231.
[18] J. Huang, G. Li, and G. Li, "RESEARCH ON SENTIMENT RECOGNITION IN ONLINE SHOPPING REVIEWS USING AN INTEGRATED BERT AND BILSTM APPROACH," 2024. doi: https://doi.org/10.1117/12.3035076.
[19] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesty, and A. Arifianto, "SENTIMENT ANALYSIS ON A LARGE INDONESIAN PRODUCT REVIEW DATASET," 2024. doi: https://doi.org/10.20473/jisebi.10.1.167-178.
[20] A. Kumar, N. K. Gupta, and A. Shrivastava, "A CRITICAL REVIEW ON SENTIMENT ANALYSIS BASED ON DEEP LEARNING TECHNIQUES," 2024. doi: https://doi.org/10.36948/ijfmr.2024.v06i05.28572.
[21] Z. Anwar, H. Afzal, T. Al-Shehari, M. Al-Razgan, T. Alfakih, and R. Nawaz, "MINING THE OPINIONS OF SOFTWARE DEVELOPERS FOR IMPROVED PROJECT INSIGHTS: HARNESSING THE POWER OF TRANSFER LEARNING," 2024. doi: https://doi.org/10.1109/ACCESS.2024.3397211.
[22] W. She, "A REVIEW OF DEEP LEARNING-BASED TEXT SENTIMENT ANALYSIS RESEARCH," 2024. doi: https://doi.org/10.54254/2755-2721/32/20230204.
[23] B. He, R. Zhao, L. Wang, and Y. Yang, "ASPECT-BASED SENTIMENT ANALYSIS BASED ON FEATURE EXTRACTION AND ATTENTION," 2024. doi: https://doi.org/10.54097/da7qvd31.
[24] A. G. Ganie and S. Dadvandipour, "EXPLORING THE IMPACT OF INFORMAL LANGUAGE ON SENTIMENT ANALYSIS MODELS FOR SOCIAL MEDIA TEXT USING CONVOLUTIONAL NEURAL NETWORKS," *Multidiszciplináris Tudományok*, 2023. doi: https://doi.org/10.35925/j.multi.2023.1.17.
[25] M. Aeri and A. Professor, "ANALYSIS OF PRODUCT REVIEWS BASED ON SENTIMENT ANALYSIS BY USING MACHINE LEARNING AND DEEP LEARNING," 2023. doi: https://doi.org/10.48047/NQ.2022.20.3.NQ22956.
[26] R. Das and T. D. Singh, "MULTIMODAL SENTIMENT ANALYSIS: A SURVEY OF METHODS, TRENDS, AND CHALLENGES," 2023. doi: https://doi.org/10.1145/3586075.
[27] S. M. Hashemi, R. M. Botez, and G. Ghazi, "BIDIRECTIONAL LONG SHORT-TERM MEMORY DEVELOPMENT FOR AIRCRAFT TRAJECTORY PREDICTION APPLICATIONS TO THE UAS-S4 EHÉCATL," 2024. doi: https://doi.org/10.3390/aerospace11080625.
[28] H. P. Pandit, "REGULAR PAPER COMPARATIVE ANALYSIS OF DEEP LEARNING MODELS FOR SENTIMENT ANALYSIS ON IMDB REVIEWS," 2024. doi: https://doi.org/10.52783/jes.1345
[29] R. A. Hameed, W. J. Abed, and A. T. Sadiq, "EVALUATION OF HOTEL PERFORMANCE WITH SENTIMENT ANALYSIS BY DEEP LEARNING TECHNIQUES," 2023. doi: https://doi.org/10.3991/ijim.v17i09.38755.
[30] S. Goswami and A. Kumar, "BILSTM_SAE:A HYBRID DEEP LEARNING FRAMEWORK FOR PREDICTIVE DATA ANALYTICS SYSTEM IN TRAFFIC MODELING," 2023. doi: https://doi.org/10.21203/rs.3.rs-2422617/v1.
[31] A. He and M. Abisado, "TEXT SENTIMENT ANALYSIS OF DOUBAN FILM SHORT COMMENTS BASED ON BERT-CNN-BILSTM-ATT MODEL," *Ieee Access*, 2024. doi: https://doi.org/10.1109/ACCESS.2024.3381515.
[32] N. Chen, Y. Sun, and Y. Yan, "SENTIMENT ANALYSIS AND RESEARCH BASED ON TWO-CHANNEL PARALLEL HYBRID NEURAL NETWORK MODEL WITH ATTENTION MECHANISM," *Iet Control Theory and Applications*, 2023. doi: https://doi.org/10.1049/cth2.12463.
[33] A. V. Ramana, "NONLINEAR DEEP LEARNING FRAMEWORK FOR PRECISE SENTIMENT CLASSIFICATION IN TEXTUAL DATA," *CANA*, 2024. doi: https://doi.org/10.52783/cana.v32.2345.
[34] X. Li, L. Chen, B. Chen, and X. Ge, "BERT-BILSTM-ATTENTION MODEL FOR SENTIMENT ANALYSIS ON CHINESE STOCK REVIEWS," 2024. doi: https://doi.org/10.21203/rs.3.rs-4023113/v1.
[35] M. R. Alifi, D. C. Utomo Lieharyani, B. P. Sudimulya, and M. R. Maulidhan, "IMPLEMENTATION OF INDONLU PRE-TRAINED MODEL FOR ASPECT-BASED SENTIMENT ANALYSIS OF INDONESIAN STOCK NEWS," *Jurnal Teknik Informatika*, 2023. doi: https://doi.org/10.15408/jti.v16i2.33791.
[36] H. S. Wicaksana, R. Kusumaningrum, and R. Gernowo, "DETERMINING COMMUNITY HAPPINESS INDEX WITH TRANSFORMERS AND ATTENTION-BASED DEEP LEARNING," *Iaes International Journal of Artificial Intelligence (Ij-Ai)*, 2024. doi: https://doi.org/10.11591/ijai.v13.i2.pp1753-1761.
[37] X. Zhao, "EXPLORING THE PERFORMANCE OF THE CNN-LSTM MODEL IN STOCK PREDICTION," *Highlights in Business Economics and Management*, 2024. doi: https://doi.org/10.54097/x9m1cm10.
[38] H. WANG and S. Zhang, "RESEARCH ON THE APPLICATION OF IMPROVED BERT-DPCNN MODEL IN CHINESE NEWS TEXT CLASSIFICATION," *Concurr. Comput.*, 2024. doi: https://doi.org/10.1002/cpe.8338.

[39] N. Sureja, N. Chaudhari, P. Patel, J. Bhatt, T. Desai, and V. Parikh, "HYPER-TUNED SWARM INTELLIGENCE MACHINE LEARNING-BASED SENTIMENT ANALYSIS OF SOCIAL MEDIA," *Engineering Technology & Applied Science Research*, 2024. doi: https://doi.org/10.48084/etasr.7818.

[40] F. Liu, C. Yuan, H. Chen, and F. Yang, "PREDICTION OF LINEAR B-CELL EPITOPES BASED ON PROTEIN SEQUENCE FEATURES AND BERT EMBEDDINGS," *Sci. Rep.*, 2024. doi: https://doi.org/10.1038/s41598-024-53028-w.

[41] B. Zeng, "PARTICLE SWARM OPTIMIZATION-BASED NLP METHODS FOR OPTIMIZING AUTOMATIC DOCUMENT CLASSIFICATION AND RETRIEVAL," *PLoS One*, 2025. doi: https://doi.org/10.1371/journal.pone.0325851.

[42] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "PLANT-DISEASE RELATION MODEL THROUGH BERT-BILSTM-CRF APPROACH," *Indonesian Journal of Electrical Engineering and Informatics (Ijeei)*, 2024. doi: https://doi.org/10.52549/ijeei.v12i1.5154.

[43] F. Wang and Z. Liu, "POWER LOAD PREDICTION BASED ON IGWO-BILSTM NETWORK," *Math. Probl. Eng.*, 2023. doi: https://doi.org/10.1155/2023/8996138.

[44] C. Wang and J. Qiao, "CONSTRUCTION PROJECT COST PREDICTION METHOD BASED ON IMPROVED BILSTM," *Applied Sciences*, 2024. doi: https://doi.org/10.3390/app14030978.

[45] W. Sun, Y. Wang, X. You, D. Zhang, J. Zhang, and X. Zhao, "OPTIMIZATION OF VARIATIONAL MODE DECOMPOSITION-CONVOLUTIONAL NEURAL NETWORK-BIDIRECTIONAL LONG SHORT TERM MEMORY ROLLING BEARING FAULT DIAGNOSIS MODEL BASED ON IMPROVED DUNG BEETLE OPTIMIZER ALGORITHM," *Lubricants*, 2024. doi: https://doi.org/10.3390/lubricants12070239.

[46] Y. Yue, Y. Peng, and D. Wang, "DEEP LEARNING SHORT TEXT SENTIMENT ANALYSIS BASED ON IMPROVED PARTICLE SWARM OPTIMIZATION," *Electronics (Basel).*, 2023. doi: https://doi.org/10.3390/electronics12194119.

[47] L. Wang, H. Wang, T. Lu, and C. Wang, "REACTIVE POWER OUTPUT MODELING OF SYNCHRONOUS CONDENSER IN UHVDC CONVERTER STATION BASED ON INTERLACED SUPERPOSITION CNN-BILSTM," *International Journal of Computational Intelligence Systems*, 2023. doi: https://doi.org/10.1007/s44196-023-00363-x.

[48] W. Ma, Y. Lei, B. Yang, and P. Guo, "A MULTI-MECHANISM FUSION METHOD FOR ROBUST STATE OF CHARGE ESTIMATION VIA BIDIRECTIONAL LONG SHORT-TERM MEMORY MODEL WITH MIXTURE KERNEL MEAN P-POWER ERROR LOSS OPTIMIZED BY GOLDEN JACKAL OPTIMIZATION ALGORITHM," *J. Electrochem. Soc.*, 2024. doi: https://doi.org/10.1149/1945-7111/ad7c7f.

[49] C. Shi, A. K. Chiu, and H. Xu, "EVALUATING DESIGNS FOR HYPERPARAMETER TUNING IN DEEP NEURAL NETWORKS," *The New England Journal of Statistics in Data Science*, 2023. doi: https://doi.org/10.51387/23-NEJSDS26.

[50] D. Qian, "ANALYSIS OF THE HYPERPARAMETER SELECTION IN MACHINE LEARNING," *Applied and Computational Engineering*, 2024. doi: https://doi.org/10.54254/2755-2721/104/20241140.

[51] R. Malhotra and M. Cherukuri, "A SYSTEMATIC REVIEW OF HYPERPARAMETER TUNING TECHNIQUES FOR SOFTWARE QUALITY PREDICTION MODELS," *Intelligent Data Analysis*, 2024. doi: https://doi.org/10.3233/IDA-230653.

[52] C. Xu, P. Coen-Pirani, and X. Jiang, "EMPIRICAL STUDY OF OVERFITTING IN DEEP LEARNING FOR PREDICTING BREAST CANCER METASTASIS," *Cancers (Basel).*, 2023. doi: https://doi.org/10.3390/cancers15071969.

[53] J. A. Ilemobayo *et al.*, "HYPERPARAMETER TUNING IN MACHINE LEARNING: A COMPREHENSIVE REVIEW," *Journal of Engineering Research and Reports*, 2024. doi: https://doi.org/10.9734/jerr/2024/v26i61188.

[54] X. Chang, S. Yang, S. Li, and X. Gu, "ROLLING ELEMENT BEARING FAULT DIAGNOSIS BASED ON MULTI-OBJECTIVE OPTIMIZED DEEP AUTO-ENCODER," *Meas. Sci. Technol.*, 2024. doi: https://doi.org/10.1088/1361-6501/ad5460.

[55] A. Simanjuntak *et al.*, "RESEARCH AND ANALYSIS OF INDOBERT HYPERPARAMETER TUNING IN FAKE NEWS DETECTION," *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (Jnteti)*, 2024. doi: https://doi.org/10.22146/jnteti.v13i1.8532.