AIA62027379_R1 (1).pdf

By Turnitin LLC

WORD COUNT	8407	TIME SUBMITTED	24-JUL-2026 05:32PM
		PAPER ID	123145621

Received: 25 August 2025 | Revised: 29 January 2026 | Accepted: 23 June 2026 | Published online: 20 July 2026

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62027379

RESEARCH ARTICLE

Deep Learning Approaches for Hate Speech Detection in Resource-Constrained Social Media Texts: A Comparative Evaluation

Yosia Immanuel Bastian[1], Aditiya Hermawan[1,*], Ardiane Rossi Kurniawan Maranto[2], Benny Daniawan[2] and Junaedi Junaedi[2]

[1] *Department of Informatics Engineering, Buddhi Dharma University, Indonesia*

[2] *Department of Information Systems, Buddhi Dharma University, Indonesia*

Abstract: Hate speech that occurred in social media platforms presents critical safety challenges, especially in language environments with minimal source where text is often written in an informal form, abbreviated, and context dependent. This paper further provides empirical evidence regarding the ability of the BERT (Bidirectional Encoder Representations from Transformers) model in detecting hate speech in a rough text environment and comparing directly with other deep learning techniques.Three Bidirectional Long Short-Term Memory (BiLSTM) variants using Word2Vec, FastText, and TF-IDF representations were compared with BERT, which employs contextual language representations. The data is separated by using a train-test ratio of 80:20, and the performance is evaluated according to their accuracy, precision, recall, F1-score, and area under the curve (AUC). The result shows that BERT outperforms all variations of BiLSTM by achieving an accuracy of 87.18%, an F1-score of 87.08%, and an AUC matrix of 0.9244. According to the efficiency matters, BiLSTM and Term Frequency–Inverse Document Frequency (TF-IDF) are the most ineffective for their uneven classification distribution, while BiLSTM with Word2Vec and FastText shows moderate effectivity. This finding conclusively demonstrates the benefit of transformer-based models to capture the subtleties of language with noisy textual data, which is commonly found in low-resource settings. To conclude the above discussion, this finding suggests that BERT has great potential as a multilingual content moderation tool to be applied in informal and unstructured digital environments.

Keywords: BERT, deep learning, hate speech detection, low-resource language, BiLSTM

1. Introduction

Working and living patterns of people have undergone huge changes under the influence of the information age and communication, which has now spread widely across mediums including social media with its sites such as Facebook, Twitter, and Instagram; posts that allow homeowners everywhere to share ideas instantly online, or media exposure, have made relatively ordinary characters into public figures virtually overnight. As an open medium, this kind of creation is naturally accompanied by impediments. The biggest problem it faces is hate speech. That kind of language is used to undermine one or more groups of people by race, religion, ethnic background, gender identity, or sexual orientation [1, 2]. Social media has now been upgraded to an issue of concern for the government of Indonesia, whose citizens use it heavily. As with the larger public, social media carries the threat of social order compromised by inter-ethnic tensions [3].

On social platforms, hate speech frequently surfaces in terse and Schadenfreude-style apathetic comments. In such a casual, situational idiom, automatic discrimination methods will find themselves in an impossible bind. Formal and context-free rating methods have been twisted by this kind of text. Thus, hate speech detection is a complicated machine learning and natural language processing (NLP) problem, requiring models to make inferences about the contextual meaning in text [4–6]. The main problem is how to design an effective classifier, especially for languages with rich configuration structure and part-of-speech context, such as Indonesian [7, 8].

BiLSTM and BERT with a Bidirectional Encoder (BiLSTM-BERT) node are two of the most popular new machine learning architectures following NLP's bidirectional approach. This means that the BiLSTM structure, unlike models such as Word2Vec and GloVe, incorporates information from both before and after the current location in an input sequence direction (forward and backward). Its error rate on a document classification task was shown to be less than half that of a standard one-direction long short-term memory (LSTM) machine. However, one of its major problems is unsuitability for cross-document tasks [9, 10]. In contrast, BERT has consistently demonstrated leading performance in NLP research, including text classification using various vocabularies and conventions, such as the General Inquirer, sentiment analysis,

***Corresponding author:** Aditiya Hermawan, Department of Informatics Engineering, Buddhi Dharma University, Indonesia. Email: aditiya.hermawan@ubd.ac.id

and the analysis of Thorndike's word corpus containing adverbs of negation. BERT's main advantage is the ability to resolve lexical ambiguity in a highly accurate manner: polysemy and complex syntactic structures are both solidly processed [11–13]. This model is effective in analyzing fine-grained word-level semantics and extracting cross-sentence association. As a transformer-based model, BERT can capture word-level semantics and automatically extract high-level features from multiple representation layers [14, 15]. This matter produces short-term effects in words or phrases in a sentence along with the long-term effects from each and every word in the documents [16, 17]. Even though BERT has been adopted for hate speech detection, including in studies using Indonesian datasets and domain-adapted variants such as IndoBERTweet, existing research has focused more on achieving higher predictive performance than on systematically analyzing how contextual and non-contextual modeling paradigms perform under identical experimental conditions. Specifically, there is less comparative study to evaluate BERT and BiLSTM architectures side by side, using the same actual Indonesian social media data, the same preprocessing pipeline, identical data splits, and consistent evaluation matrix, especially when several embedding word strategies are used [18].

Revolutionary technological developments, such as advances in word representation techniques including Word2Vec, FastText, and TF-IDF, fundamentally have constructive acts in model performance, not to put aside the importance of classification algorithm developments. On the other hand, Word2Vec and FastText change words into distributed vector representations, where the continuous semantic space captures the semantic relationships among different words. In contrast, TF-IDF gives statistical weights in terms of any documents according to the relative importance of other terms in the same document [19]. By combining with a deep learning algorithm, this embedding technique has demonstrated enhancement in detecting hate speeches [20]. However, comprehensive studies that systematically compare the effects of several embedding strategies impacts toward BiLSTM performance and explicitly compare that result with transformer-based models that inherently learn contextual embedding are hardly ever available for informal and noisy Indonesian social media texts.

To address this gap, this study adopts a comparative and controlled diagnostic research design. This paper further provides two main contributions: first, by doing systematic evaluation along with multilanguage baseline of the BERT model and three variants of BiLSTM that use different embedding strategies (Word2Vec, FastText, and TF-IDF), which are all trained and tested using the same dataset, preprocessing pipeline, data splits, and evaluation matrix, and second, by isolating the contextual representation effects versus the non-contextual effects. This study provides empirical insights into why transformer-based models tend to be more robust in low-resource, informal, and context-dependent linguistic environments. All phases of research follow the standardized experiment design to ensure methodological rigor and reproducibility. Furthermore, this study aims to support the decision-making in choosing and implementing a model for the moderation system of Indonesian hate speech and other minimum resource language contexts.

2. Literature Review

Hate speech detection is now becoming a hot topic for research in the domain of NLP. This is driven by the massive spread of user-generated content in social media platforms. The prior studies for this area mostly relied on traditional machine learning techniques like support vector machines (SVMs), Naïve Bayes, and logistic regression for automatic text classification tasks [1, 3, 8]. Even though their models are considered successful for these tasks, they are likely to rely on the features, which means they rely on the manually factored variables and bag-of-words representation, which is inherently unable to capture the semantic complexity or contextual nuances as the characteristic of informal online discourses. Hence, these classic models exhibit limited robustness when applied to noisy social media texts that include sarcasm, abbreviations, and implicit hate expressions.

To address this limitation, upcoming research introduces a sequential and hybrid deep learning architecture that utilizes distributed word representation and temporal dependencies. For context, Riyadi et al. [2] employed a Convolutional Neural Network–Recurrent Neural Network (CNN–RNN) hybrid model for Indonesian hate speech detection and reported notable improvements over traditional classifiers. Their improvement compared to traditional approaches is noticeable. In a similar fashion, Dwitama et al. [9] turned to BiLSTM for Indonesian Twitter data and found good results. At the same time, Ghozali et al. [21] showed through feature expansion strategies that BiLSTM performance could be improved even further. Nevertheless, even with this kind of development, sequential RNN-based architectures such as BiLSTM are still limited in capturing long-range dependencies and subtle pragmatic cues, especially for hate speeches that are implicit and sarcastic. This challenge is enhanced in a super noisy environment that involves code-mixing, slang words, and common informal expressions that are commonly found in social media. Referring to the recent study by Saputra and Sibaroni [17], adding auxiliary features (such as statement length) in a BiLSTM-based model produces a better score in multi-label political discourse classification. Furthermore the development, NLP research has undergone a paradigm shift toward transformer-based architectures, which model global contextual relationships through a self-attention mechanism. Among these models, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a dominant framework due to its bidirectional contextual encoding, strong handling of polysemy and syntactic ambiguity, and state-of-the-art performance across diverse NLP tasks, including sentiment analysis, text classification, and question answering [11, 13, 15]. There is also additional evidence of the superiority of transformer-based techniques in hate speech classification from Mutanga et al. [22], Saleh et al. [23], and Kusuma and Chowanda [24]. In the context of Indonesian, we need another example that shows using IndoBERTweet and BiLSTM can reach 93.7% for classification accuracy. Other studies enhanced the BERT model for low-resource scenarios, for example, the Triple-Compressed BERT of Peng and Zhao [12], or transfer learning methodologies investigated by Devianty [16], which is tested to be able to handle the classification task well even with a small training dataset. However, despite the promising findings, systematic empirical comparison between transformer-based models and sequential deep learning architectures under identical experiment condition are still rare, especially for languages with limited resources like Indonesian.

Word representation is the main research focus for other researchers. Especially now, there are plenty of effective ways to transform textual data into digital features that can be used for classification tasks, such as Word2Vec, FastText, and TF-IDF [19, 20, 25]. This approach simply appears in a lot of studies, such as sentiment classification and hateful speech research, but it has been found to be weak. Context, sarcasm, and various verbal idioms so common on social media go unrepresented [26–28]. Recent reviews, such as that by Asudani et al., have put extra

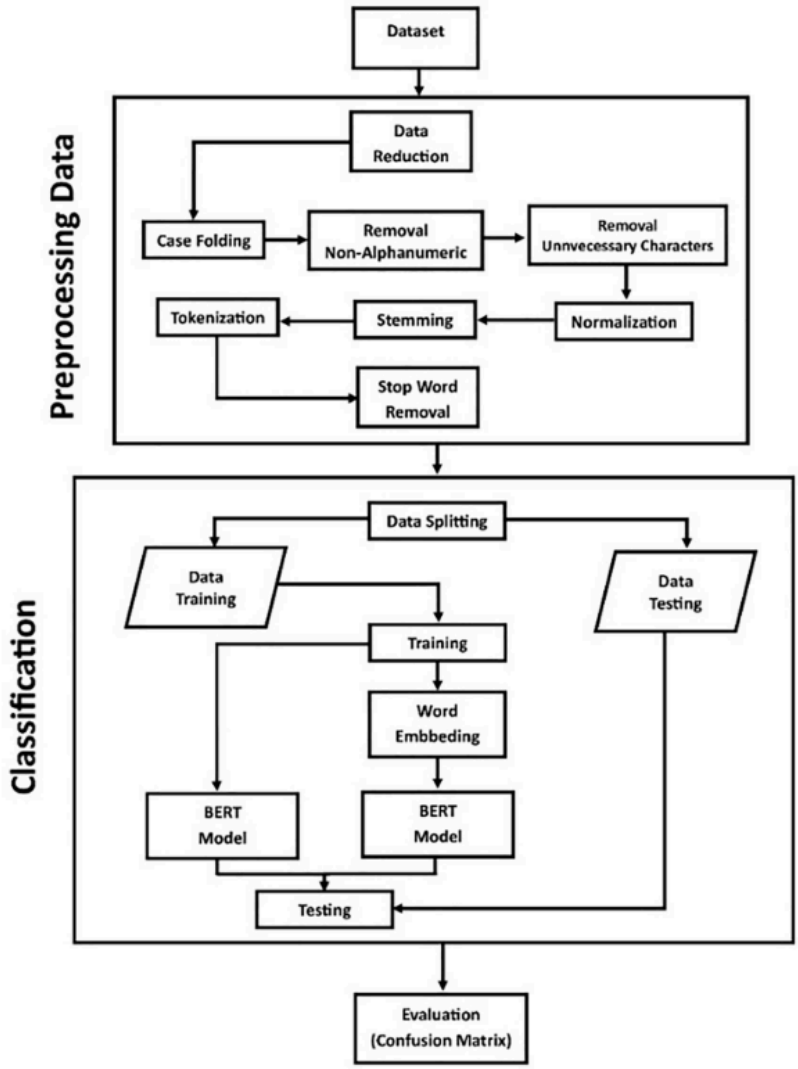
emphasis on contextual embeddings over static ones. In this age of informal social media texts, it becomes more apparent that non-contextual models are less useful [25].

Even with several progresses, only a few of empirical studies directly compare contextual models with hybrid models with the same embedding strategy in an identical experimental framework such as sequential BiLSTM, especially for languages with limited resources such as Indonesian [18]. In the current benchmark studies, the focus still lies on the sentiment instead of hate speech. For the current state, this field is limited to classic machine learning [1, 3, 8], hybrid deep learning model [2, 9], or transformer-based evaluation only [22, 23]. Given the lack of comprehensive comparative studies examining the results of these three perspectives (which are transformer-based model, sequential deep learning model, mixed model) in noisy, informal, transient online discourse, it is hard to find out how effective the applied approach is if it is applied in other languages and cultures. It is important to fill in the discrepancies: a good and compatible structured hate speech detection system for minoritized languages, which is aimed to be possible by such research.

3. Research Methodology

This research uses quantitative experimental methodology to systematically analyze and compare several deep learning approaches in detecting hate speech in Indonesian social media contents. This experiment is done using a Twitter corpus, which reflects the characteristics of actual online discourses that are informal, noisy, and context dependent. Besides accuracy comparison, explicit methodology design is aimed to highlight the comparative distribution from several modeling paradigms in controlled and unified experimental settings. The overall experimental procedure follows the structured life cycle, as pictured in Figure 1 (Research Methodology Framework), including data collection, preprocessing text, word representation, model training, and performance evaluation, followed by the selection of the best-performing model to be integrated into the web-based prototype. The novelty of this methodology lies in the design as a comparative evaluation cycle, which simultaneously tests the transformer-based model, sequential deep learning model, and embedding-driven variants in the same experiment framework. By controlling the stages of preprocessing, data splitting, training

Figure 1
Research methodology framework



configuration, and evaluation matrix, this study enables fair and interpretable comparison among models that are often evaluated separately in previous work.

3.1. Data collection and data labeling

The data used in this study is derived from Twitter social media in Indonesia. The used datasets are public and developed by Ibrohim and Budi [29], which are available on the Kaggle website, which allows users to search further for this study without extra cost. The big categories in the dataset are hate speech (labeled 1) and non-hate speech (labeled 0). All entries from the previous dataset have been annotated and validated according to established hate speech labeling criteria, hence ensuring the robustness and consistency of the annotation. Utilizing the existing and validated datasets allows this study to be more focused on methodological comparison than subjective annotation, while also facilitating the reproducibility and comparability with prior research.

3.2. Data preprocessing

As the first step in transforming raw textual inputs into a form suitable for machine learning-based classification [30, 31], data preprocessing is a critical foundation item. In the present study, preprocessing was designed to remove extraneous elements and standardize input structures to ensure consistency across samples [26–28]. A standardized three-stage preprocessing pipeline is applied homogeneously across all models to eliminate the distracting effects that occurred as a result of inconsistent input preparation.

The first stage is case folding. This function changes all alphabet characters to lowercase and ensures the model interprets words lexically, such as Benci, BENCI, and Hate as equivalent. This ensures there will be no extraneous variation due to letter casting. Hence, models might use their resources to learn semantic aspects rather than differentiate words based on shallow differences.

The second stage is tokenization, a process to separate texts into word units (called tokens). By doing tokenization, longer text might be divided into smaller units; hence, the algorithm is able to do the processing and classification. The third stage of this pipeline is removing stopwords, which includes eliminating irrelevant common words for analysis like "yang," "dari," and "dan." This stage aims to reduce the noise in the data and make the model more efficient. In other words, the preprocessing stages are reflected in Table 1.

The completion of these three phases is deemed adequate for generating text that is clean and prepared for the subsequent step, which involves word representation (word embedding) and the training of a classification model utilizing the BERT and BiLSTM algorithms.

3.3. Word embedding

The conversion of textual input into a numeric vector, which is called word embedding in technical phrases [25, 32]. In this study, word embedding is explicitly used in BiLSTM mode, while BERT inherently produces contextual embedding during the encoding process; hence, there are no external embedding layers needed. The three different word representations—Word2Vec, FastText, and TF-IDF—are integrated into the BiLSTM architecture to examine how several embedding paradigms affect the sequential model performance. By comparing frequency-based representations (TF-IDF) with semantically aware embeddings (Word2Vec and FastText), this study isolates the contribution of embedding richness in a fixed sequential modeling framework. This design allows a clearer understanding of different performances that happen due to the embedding strategy or basic model architecture.

3.4. Modeling

After the dataset is cleared out from the preprocessing stage, the dataset is then split 80:20; 80% for model training and 20% for new model testing. This fixed ratio is applied consistently to all models to ensure the comparison and objective assessment of the generalization ability. Then BERT and BiLSTM as deep learning architectures are evaluated. These two models are designed to do binary classification to determine whether the Indonesian Twitter text is considered hate speech or not. The modeling strategy is intentionally comparative, with all experiment variables being guarded constantly, except the main modeling paradigm (contextual transformer vs. sequential recurrent architecture) and input representation strategy.

3.4.1. Model BERT

BERT is a deep learning model based on the transformer structure. Specifically, this model has been modified to understand the overall context of a word in a sentence, not by receiving input tokens in sequence like LSTM that are only able to process the token one by one [33]. As a result of this design, BERT uses "multi-head attention" so that parts of the input are able to be read simultaneously. Hence, the model is able to accurately understand the interpretation of the given words in each and every context. It provides significant convenience in classifying hate speech and takes explicit cues or sentiment from context. A BERT architecture-based model (i.e., BERT with a case-mix version for

Table 1
Data processing stages

Tweet	Stages	Results
#2019GantiPresiden scra konstitusional hak setiap rakyat woiii!!. Gini nih kalo planga plongo tapi kelakuan otoriter	Case folding	#2019gantipresiden scra konstitusional hak setiap rakyat woiii!!. gini nih kalo planga plongo tapi
	Tokenizing	['2019', 'ganti', 'presiden', 'konstitusional', 'hak', 'rakyat', 'woiii', 'nih', 'plonga', 'plongo', 'laku', 'otoriter']
	Stopword removal	#2019GantiPresiden scra konstitusional hak rakyat woiii!!. Gini nih kalo planga plongo kelakuan otoriter

multilanguage support) is used in this study. This model is highly suitable for social media content analysis from languages with limited resources, since this model has been pre-trained to understand several multilanguage contexts, including Indonesian, which often contains a lot of noise and abbreviations.

The classification process starts with the WordPiece tokenizer, which is used to tokenize every tweet. In this stage, special tokens [CLS] and [SEP] are positioned at the start and end of the raw text. The next step is to rearrange the sequence into a fixed-length numerical encoding that incorporates the embedding token, segment, and position. Every token winds up as an array of 768 numbers. The embedding is then forwarded through 12 layers of transformer encoders. This structure allows models to build complex contextual representations by considering the syntax of sentences and latent relationships between words across the whole sequence. The [CLS] embedding is formed by adding together the output of 12 layers simply by their weights, and then the result will be treated as an average representation for further analysis. To extract the final layer, the token corresponds to the embedding [CLS] vector. As demonstrated in Figure 2, embedding [CLS] is included in a dense layer (fully connected) with a softmax activation function, hence doing binary classification to provide a label to every tweet as hate speech or non-hate speech. The model is trained by using loss cross-entropy function and optimized with AdamW. The training is done with a 2e-5 learning rate and a batch size of 32. The model is trained with a total of 10 epochs. Assessment techniques include conventional metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). In effect, this combination of architectural components and exploitation of BERT's contextual learning capabilities enables the model to discern hate speech even in subtle cases such as hidden allusions, slurs wrapped around humor, or languages so stylized that one would not recognize them for sarcasm.

3.4.2. Model BiLSTM

In this study, we use BiLSTM as a sequential learning base and test how well the transformer-based BERT model performs in detecting hate speech. Unlike the usual LSTM, the text input in BiLSTM is processed in two directions at once: forward (from left to right) and backward (from right to left). This bidirectional model allows to capture contextual clues, whether preceding or ensuing in the sentence, thereby enhancing its ability to bring the chunks of semantic information. This is relevant for social media text, which is usually informal, abbreviated, and very context dependent. Regarding this matter, BiLSTM is used to read widely and deeply into the local knowledge since it is able to conclude the meaning of each word in its surrounding context. So, BiLSTM covers a wider information range compared to LSTM, which allows it to differentiate linguistic nuance even if it is not an idiom. Starting with the input sentence that has been preprocessed and converted into a word vector by using one of these three techniques, Word2Vec, FastText, or TF-IDF, the text classification process with BiLSTM can be further reflected in Figure 3. Each

Figure 2
Text classification process architecture using BERT

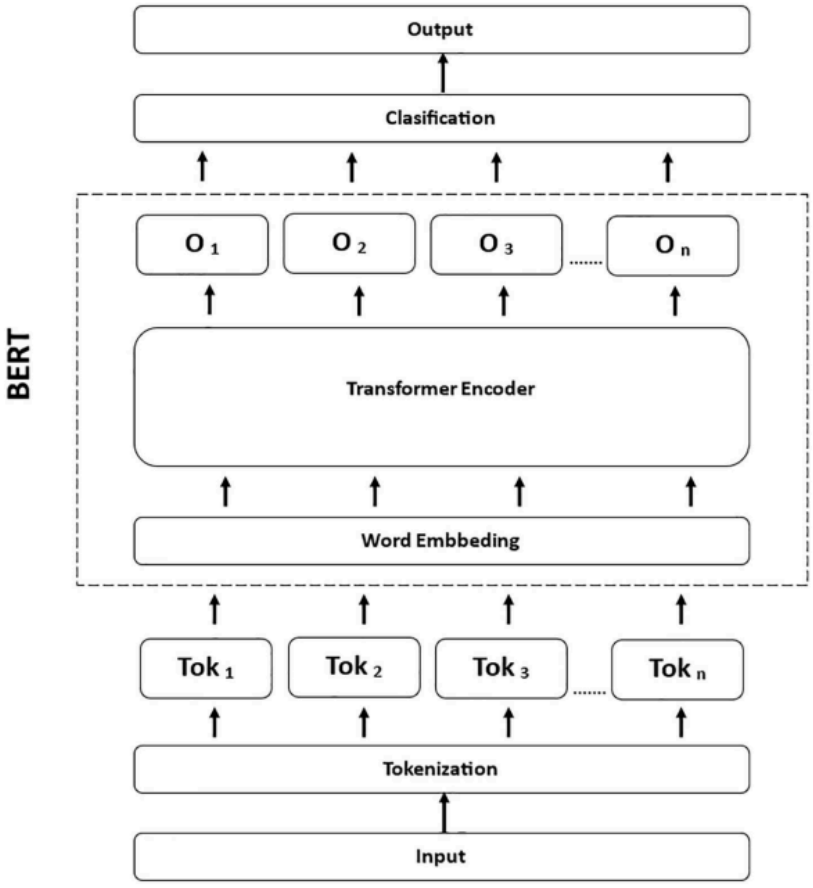
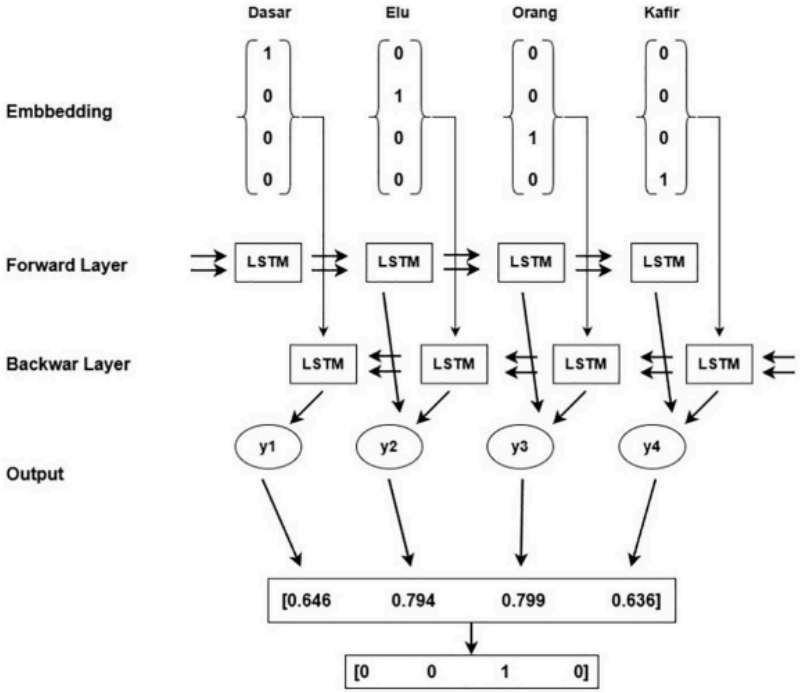


Figure 3
BiLSTM modeling architecture for text classification



of these three methods serves as an input representation mode and is chosen to examine the influence of different coding strategy features on the model performance. The embedded tokens are further inputted into two parallel rows of LSTM cells: one row from left to right and the other from right to left. Each cell outputs the current state for all tokens, and after combining by using concatenation, the forward hidden states and backward ones are able to provide a more comprehensive word representation. To achieve a fixed-length vector for classification, the token-level output is then averaged (i.e., global average pooling) or chosen as a certain token representation (i.e., middle or final token) and then forwarded to the classification layer. The model input layer utilizes a standard classification design, consisting of a densely connected layer with a softmax activation function. This design further provides a probability score for our binary classification task to distinguish between hate speech (label 1) and non-hate speech (label 0). The model is trained by using a cross-entropy categorical loss function, combined with the Adam optimizer. For the consistency of every embedding combination, it is necessary to maintain a similar experimental setting in each testing. Fixed training configuration is used: batch size of 32 and 10 iteration training. This architecture is deliberately chosen so that the comparison between BERT and BiLSTM models can be done as fairly and closely as possible. By fixing all variables, except the main sequential modeling method and embedding technique for inputs (whether it is contextual or not), we are able to separate between contextual model effect and non-contextual in classification performance.

Accordingly, the study seeks to achieve two objectives: subsidiarily (1), to depend on BiLSTM as a sequential learning model for detecting hate speech and formalizing its intrinsic capabilities, and mainly (2), to quantify the contribution that different embedding strategies make within a minimized-resource, noisy environment in an informal text context.

3.5. Evaluation

This study chooses the confusion matrix as the evaluation strategy that the model follows. The dataset is divided into two parts: 80% for training and the other used exclusively as test data [34]. Specifically, for each instance in the test set of a model trained on the training subset alone, the predicted label is compared to the corresponding ground-truth label for constructing the confusion matrix; that is, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) entries follow.

The use of several complementary matrices ensures that the evaluation is not only limited to overall accuracy but also includes the ability to classify the class and model robustness. These matters are important and crucial to consider in detecting hate speech in real-life social media.

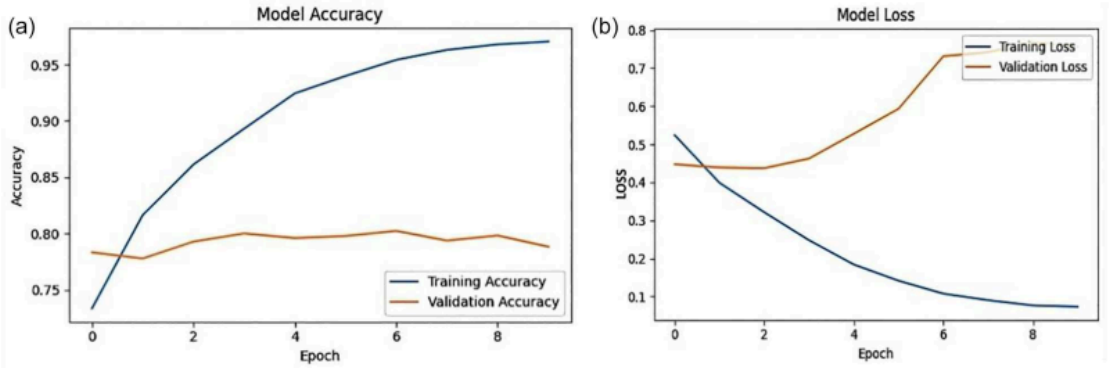
4. Results and Discussion

Four candidate model classifications: BiLSTM-Word2Vec, BiLSTM-FastText, BiLSTM-TF-IDF, and BERT. In this paper, a dataset is divided into 80% for training and 20% for testing. This allows us, with any of these models, to get hands-on experience on how well they are trained. The BiLSTM configurations were then trained batch by batch. The batch size was 32, and each data point was added to it as one epoch went by. The purpose of this study is to evaluate the performance of each model in classifying hate speech by using a standardized matrix.

4.1. BiLSTM model evaluation—Word2Vec

Figure 4(a) shows the accuracy trend for 10-epoch training. The accuracy is increasing steadily and exceeds 95% in the 9th epoch. However, the accuracy of validation peaks between 78%

Figure 4
Graph of accuracy model (a) and loss model (b) BiLSTM-Word2Vec



80%, with minimal fluctuation toward the end of the epoch. The widening gap between the accuracy of training and validation indicates that there is overfitting occurs, where the model performs well on the training data but shows limited generalization ability on the unseen samples. In Figure 4(b), training and validation losses are depicted as corollaries. The training loss continued to decrease throughout, which was consistent with the promise that it couldn't possibly be learning from the training set. After the third epoch, however, validation loss rose, with a peak around the seventh epoch. In validation, where the model no longer finds refutations easy to come by, the rising trend of loss corroborates overfitting even more. Basically, this indicates that where the unnatural patterns turn more extreme, complex, or difficult for an overfitted model to reconstruct accurately from training data, the probability of being wrong increases.

The model classification effectivity is measured by the confusion matrix. The assessment shows an accuracy level of 80.23%, which indicates that most of the models are successful. With a precision level of 80.4%, this model is suitable to minimize false alarms. A recall level of 80.0% means that the model is accurate in a lot of cases and successfully identifies the real hate speech. F1-score, which reflects the compromise between both values, comes at 80.2%. The above matters indicate a good operational balance. Moreover, this model has an AUC value of (has also achieved an excellent separation between the two types of samples—hate speech from non-hate speech. AUC is approaching 1, its level with clear hues of blue and red—which means that in this model, higher likelihood ratings for positive examples are assigned consistently across threshold values. The idea cannot be shaken by changes simply because its machine intelligence has one single-pointed goal.

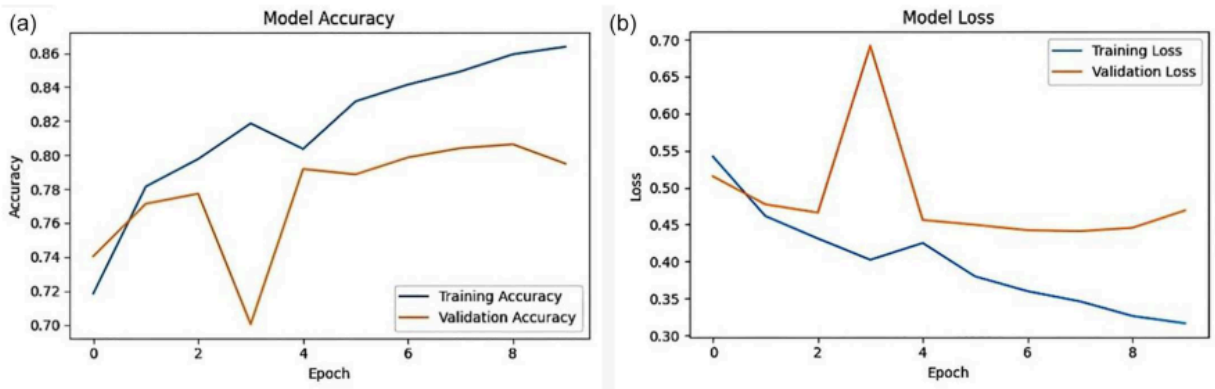
4.2. BiLSTM model evaluation—FastText3

As shown in Figure 5(a), by the 9th epoch, the training accuracy gradually increases to more than 86%. On the other hand, validation accuracy fluctuates and remains stubbornly stuck just above 80%, indicating a mismatch between predictions of this quantity (test set data) and what is expected therefrom.

In contrast, Figure 5(b) shows that the train loss continues to decrease slowly while validation losses are briefly high peaks followed by stabilization, a situation which indicates arbitrary variability between learning patterns on validation set points and training loss trends over time. In general terms, this case illustrates how one might still go about improving model validation stability to stave off any possible further overfitting down the line. The BiLSTM–FastText model achieved an overall accuracy of 80.6% as evaluated using confusion matrix classification performance, an adequate prediction capability. High-precision models are shown by the fact that 79.6% of the predictions are positively correct, while the incorrect predictions are minor, with relatively low numbers.

With the recall value of 82.3%, we view this model as factually capturing most of the hate speech cases correctly. The F1-score of 81.1% reflects the maintained balance between two aspects, precision and recall. This indicates that in class

Figure 5
Graph of accuracy model (a) and loss model (b) BiLSTM-FastText



prediction forecasting, models are in an effectively stable and effect domain. The AUC for the model is 0.88, which indicates its effectiveness in differentiating hate speech from non-hate speech. An AUC of nearly 1 indicates that models always provide a high level of trust in TP at several threshold values, as a differentiated attribute. In this matter, the difference to BiLSTM-FastText as the on-par rival of transformers is minimal, even though the approach is untidy.

4.3. BiLSTM model evaluation—TF-IDF

Figure 6(a) indicates that both the training and validation accuracies remain stagnant at approximately 49–50% over 10 epochs, without any significant increase. This indicates that models do not capture the basic patterns in the data effectively. As confirmed in Figure 6(b), both the training loss and validation are almost kept stable along the epoch, indicating that since the beginning, the process is unable to break out. This situation stands for major underfitting, where even the training data cannot be adequately fitted by the model, let alone generalizing to other information.

Performance analysis shows that the overall accuracy of the BiLSTM–TF-IDF model was 50.0%, which corresponds to random guessing. A precision rate of 50.0% suggests that all positive predictions are wrong, yet recall is 100.0% since the model classified all samples as positive. But this improved recall is illusory, as the model does not fully recognize the negative class at all. Hence, the resulting F1-score of 66.7% emphasizes the imbalance that has arisen from the inability to distinguish between categories.

Further indication that the BiLSTM-TF-IDM model is not better than random assignments in matters of performance is indicated by the AUC score of 0.50. This result is the concrete proof for classification tasks like ours, where standard language usage rules are used to generate an embedding model and a BiLSTM to be applied in the particular embedding. Hence, we need a more informative representation that is able to enhance the performance level.

4.4. BERT model evaluation

As shown in Figure 7, training and validation losses are illustrated as above. Within the training set, the training loss steadily drops until it becomes close to zero, where the model has trained strongly. However, the validation loss at just a few epochs onward soon increases sharply and peaks in round 10. This trend suggests overfitting, where the model learns the sample data by heart but cannot generalize to see a new sample in long-term memory.

Even so, BERT is able to provide higher qualification results compared to the BiLSTM-based architecture. The performance results indicate that BERT consistently earned the highest scores for all four performance indicators: 87.18% correct answers; 86.5% precision; 87.6% recall rate, the almost perfect score; and 87.1% F1-score of. These numbers indicate a good balance in detecting hate speech, yet minimizing both e alarms and omissions. Other than that, the model achiev n AUC score of 0.92, indicating high capability to distinguish hate speech and non-hate speech categories. A near-1 AUC score indicates that BERT consistently gives a higher probability score for TP compared to FP. Hence, this data confirms its effectivity in this study. Overall, this result reflects the superiority of transformer-based architecture compared to models that based by processing keywords and phrases in English texts with tag, informal, and noisy in social media. They further utilized the contextual embedding and attention mechanism to capture the soft semantic nuance from the context that might be lost if only looking for the overall context of a single word or phrase.

4.5. Comparison of model performance

To strengthen the comparative dimension of this study, performance evaluation can also be done at two levels: (1) comparison between internally trained baseline models (BiLSTM with different embeddings vs. BERT) and (2) comparison between our best-perform model and established benchmarks from previous research on hate speech detection in Indonesia. As well as this dual, we can easily understand if we make good modeling choices, along with how the obtained results compared with literature.

As shown in Table 2, our BERT implementation achieves the highest overall performance across all evaluation metrics. Outperforming all BiLSTM variants, our BERT implementation achieves the highest score of any comparison in Table 2 among the various methods tested.

Even though the BiLSTM models with Word2Vec and FastText embeddings achieved reasonable performance, with approximately 80% accuracy, they were unable to fully capture contextual aspects and the implicit meanings of the content. Inadequate learning and overfitting in the discriminator are the main reasons why the TF-IDF variant performs poorly.

Figure 6
Graph of accuracy model (a) and loss model (b) BiLSTM-TF-IDF

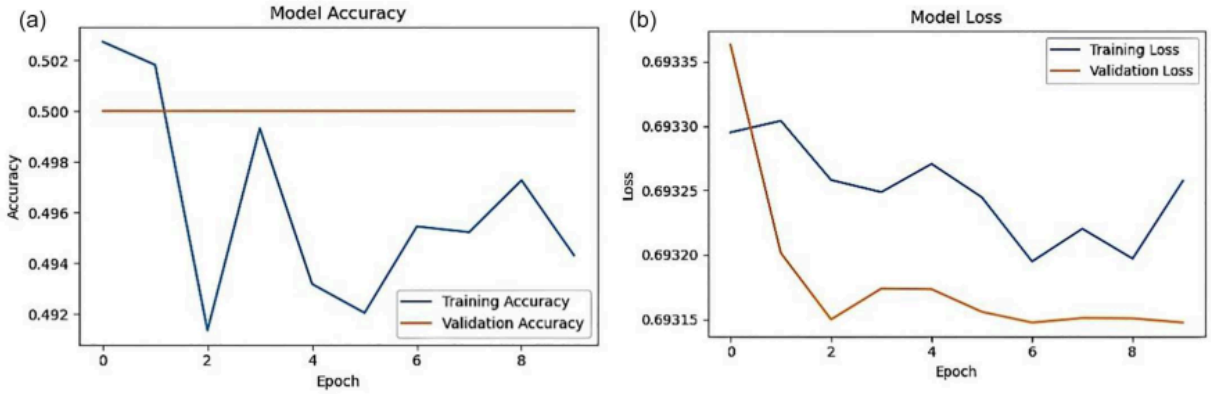


Figure 7
BERT model loss graph

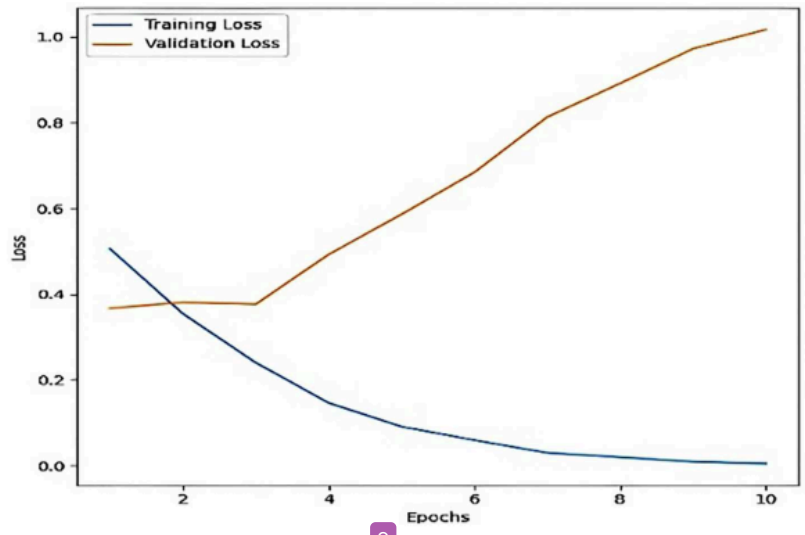


Table 2
Performance of internally trained models

Model	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
BiLSTM + Word2Vec	80.25	80.00	80.40	80.32
BiLSTM + FastText	80.6	82.30	79.60	81.07
BiLSTM + TF-IDF	50.00	100.0	50.00	66.67
BERT	87.18	87.64	86.53	87.08

In Table 3, the results show that our BERT model outperforms all methods previously applied to this dataset. But the XAI-based algorithm method can also give (correct) interpretability and precision of 83%; by contrast, traditional machine learning models such as the SVM and hybrid CNN-RNN just perform basically well themselves, getting mixed results from 74% to 86%. But BERT is always producing better prediction results: its accuracy rate is 87.18%.

This finding raises two points. First, from the point of view of internal authentication, the transformer-based method is much better than all sequential BiLSTM baseline systems presented in this article, which were created for the study—especially in terms of dealing with colloquial expressions, sarcasm, or abbreviations.

Second, our BERT results illustrate the importance of contextual embedding and self-attention mechanisms in textual features for explaining hate speech within an informal, resource-scarce linguistic environment, outperforming both the classical model and trade in external validation when compared to the previous studies' results.

Table 3
Comparison with previously reported models

Model/method	Accuracy (%)	Reference
SVM (baseline klasik)	74.88	[3] Hana et al. (2020)
Hybrid CNN-RNN	86.30	[2] Riyadi et al. (2023)
XAI-based model (XGBoost)	83.30	[35] Ibrahim et al. (2022)
BERT	87.18	This study

4.6. Discussion

According to previous studies, the BERT model performs better than BiLSTM in recognizing hate speech, especially on the Indonesian internet and Twitter as a social media platform. This matter aligns with the hypothesis in the Introduction that transformer-based models like BERT are able to do two important matters in text processing: handling long-range dependencies and in-depth language understanding. BERT accuracy of 87.18% (AUC: 0.92) for hate speech classification is in line with prior research, proving that the transformer-based model is superior in coping with natural language nuance and semantics compared to other methods [22]. In a language environment that has high-standard resources (i.e., English), transformer-based approaches including BERT indicate a strong performance in detecting hate speech, supported by large-scale pre-training corpus and richer linguistic resources [23]. According to Kusuma and Chowanda [24], the combination of BiLSTM and IndoBERTweet is able to elevate classification performance all the way up to 93.7%. It shows that for NLP in Indonesia, a model that can be sensitive to the characteristics of that particular language is highly needed. The performance gap between these results and this study's findings demonstrates that the availability of linguistic resources and domain adaptation is crucial in determining BERT effectivity in a multiple-language environment. The operated LSTM still works well even though outperformed by BERT. This result is consistent with the survey conducted by Ghozali et al. [21], who found that the use of the BiLSTM model with synonym-based

feature expansion resulted in higher accuracy. However, RNN-based models are usually unable to capture implicit context or hidden hate speech. This limitation is highlighted in the social media context with limited and noisy resources, where informal syntax, code-mixing, and creative language usage reduce the effectivity of sequential modeling. Specific domain embedding of hate speech detection has been shown time and again in the literature to be essential for accuracy. Saleh et al. find that incorporating domain-specific embedding in the model can assist in obtaining indirect hate speech ramifications. This is significant because people often use euphemisms and abbreviations to mask hate speech so as to avoid automated screening systems [23]. Explainability is crucial for hate speech detection systems, as Ibrahim et al. [35] reported. Even if BERT is considered high in accuracy, this model remains to be the "black box." To win the trust of users to the system, future development should consider the integration with Explainable AI (XAI) models, such as LIME.

One of the study developments for this research is to retrain the Indonesian BERT model to be wider and more representative, as done by IndoBERTweet [36]. Moreover, as also demonstrated by Perifanos and Goutsos [37], a multimodal approach between text input and images is a promising future direction, which may allow us to be more confident in detecting hate speech and facing the endless stream of new negative contents.

Furthermore, from this research, practical applications are expected, such as a content management system in Indonesian social media who utilizing hate speech classifier with BERT-based as the backbone to examine the content before being published. However, the application of this matter in a low-source environment requires careful validation in several linguistic and cultural contexts to determine the bias and inaccuracy [38]. Hence, further future research is needed.

As a result, this study provides empirical evidence regarding the adaptation ability of BERT-based models while shifting from a standard with a high-resourced language environment into an informal and low-resourced social media context. This might be the starting point in detecting hate speech precisely, openly, and adaptively, which is able to follow the rapid changes of language in social media.

5. Conclusion

This study compared the effectivity of the BERT model with BiLSTM in identifying hate speech in Indonesian social media, specifically Twitter. The result shows that BERT outperforms BiLSTM in all evaluation matrices with better judgment ability. With an accuracy of 87.18% for BERT, 87.80% above the closest rival, BiLSTM. The performance of BERT can be scored higher since its detailed understanding of the context of sentences through context-based representation and self-attention mechanism. Object of Paper: the main contribution is to provide comparative analysis from two well-known deep learning techniques in Indonesian NLP, which to date have not been researched adequately. This study also highlights the importance of choosing the best representation technique and giving perspective on the effectivity on transformer-based approach in classifying text by using social media data.

However, this study still has several limitations. First, data is only obtained from one source (Twitter), so the findings may not reflect the same on other platforms such as Facebook or Instagram. Second, the used BERT model is not fully adapted with informal language's structure, which is very typical in Indonesian social media (e.g., IndoBERTweet), so this matter needs to be further researched. Furthermore, a multimodal approach that combines text and visual elements needs to be developed to detect hate speech that spreads in image or memes. In conclusion, XAI techniques must be integrated to enhance the transparency, trust, and credibility of the model output that is provided to users.

Acknowledgement

Buddhi Dharma University facilitated this research by providing essential institutional and academic assistance throughout the study. The Faculty of Science and Technology contributed valuable resources, including access to research materials and computational facilities, for which the authors express their gratitude. Special thanks are extended to peers and colleagues in the departments of Information Systems and Informatics Engineering for their insightful discussions, support, and encouragement during the research journey.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available on GitHub at https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection, reference number [29].

Author Contribution Statement

Yosia Immanuel Bastian: Conceptualization, Methodology, Formal analysis, Resources, Data curation. **Aditiya Hermawan:** Software, Supervision, Project administration, Writing – original draft, Writing – review & editing, Visualization. **Ardiane Rossi Kurniawan Maranto:** Validation, Investigation. **Benny Daniawan**: Validation, Investigation. **Junaedi Junaedi:** Validation, Investigation.

References

[1] Wijaya, B., & Mawardi, V. C. (2023). Pendeteksi ujaran kebencian dada platform media sosial Twitter menggunakan support vector machine. *Jurnal Serina Sains, Teknik dan Kedokteran*, *1*(1), 11–17. https://doi.org/10.24912/jsstk.v1i1.22746

[2] Riyadi, S., Andriyani, A. D., Masyhur, A. M., Damarjati, C., & Solihin, M. I. (2023). Detection of Indonesian hate speech on Twitter using hybrid CNN-RNN. In *2023 International Conference on Information Technology and Computing*, 352–356. https://doi.org/10.1109/ICITCOM60176.2023.10442041

[3] Hana, K. M., Al Faraby, S., & Bramantoro, A. (2020). Multi-label classification of Indonesian hate speech on Twitter using support vector machines. In *2020 International Conference on Data Science and Its Applications*, 1–7. https://doi.org/10.1109/ICoDSA50139.2020.9212992

[4] Zamsuri, A., Djusar, S., & Elvira, E. (2025). Deteksi Hate Speech Pada Pemilu 20204 Menggunakan Algoritma Machine Learning. *ZONAsi: Jurnal Sistem Informasi*, *7*(1), 228–241. https://doi.org/10.31849/zn.v7i1.22049

[5] Alghifari, D. R., Edi, M., & Firmansyah, L. (2022). Implementasi bidirectional LSTM untuk analisis sentimen terhadap layanan grab Indonesia. *Jurnal Manajemen Informatika*, *12*(2), 89–99. https://doi.org/10.34010/jamika.v12i2.7764

[6] Subhi, Y. A. (2025). Klasifikasi sentimen menggunakan metode passive aggressive dengan menggunakan model bahasa BERT pada dataset kecil. *Building of Informatics, Technology and Science*, *6*(3), 1838–1847. https://doi.org/10.47065/bits.v6i3.6389

[7] Herwanto, G. B., Ningtyas, A. M., Mujiyatna, I. G., Trisna, I. N. P., & Nugraha, K. E. (2021). Hate speech detection in Indonesian Twitter using contextual embedding approach. *Indonesian Journal of Computing and Cybernetics Systems*, *15*(2), 177. https://doi.org/10.22146/ijccs.64916

[8] Rahman, O. H., Abdillah, G., & Komarudin, A. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Rekayasa Sistem Dan Teknologi Informasi*, *5*(1), 17–23. https://doi.org/10.29207/resti.v7i2.4642

[9] Dwitama, A. P. J., Fudholi, D. H., & Hidayat, S. (2023). Indonesian hate speech detection using bidirectional long short-term memory (Bi-LSTM). *Rekayasa Sistem dan Teknologi Informasi*, *7*(2), 302–309. https://doi.org/10.29207/resti.v7i2.4642

[10] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*, 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997

[11] Huang, G., & Li, Y. (2025). A novel machine reading comprehension approach through attention mechanism-based BERT model. *Journal of Circuits, Systems and Computers*, *34*(10), 2550220. https://doi.org/10.1142/S0218126625502202

[12] Peng, Z., & Zhao, Y. (2023). Triple-Compressed BERT for efficient implementation on NLP tasks. In *2023 3rd International Conference on Electronic Information Engineering and Computer Science*, 1162–1165. https://doi.org/10.1109/EIECS59936.2023.10435469

[13] Shen, J. (2025). Corpus classification algorithm based on BERT pre-trained model. *Procedia Computer Science*, *262*, 368–377. https://doi.org/10.1016/j.procs.2025.05.064

[14] Xiao, M., Yao, M., & Li, Y. (2020). Short-text intention recognition based on multi-dimensional dynamic word vectors. In *Journal of Physics: Conference Series*, *1678*(1), 012080. https://doi.org/10.1088/1742-6596/1678/1/012080

[15] Janzso, A. (2021). Disambiguating grammatical number and gender with BERT. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, 69–77.

[16] Fairuz, A. D. (2025). Transfer learning methods for hate speech detection in Bahasa Indonesia. *Journal of Information Technology and Computer Science*, *10*(1), 14–25. https://doi.org/10.25126/jitecs.2025101542

[17] Saputra, R. A., & Sibaroni, Y. (2025). Multilabel hate speech classification in Indonesian political discourse on X using combined deep learning models with considering sentence length. *Jurnal Ilmu Komputer dan Informasi*, *18*(1), 113–125. https://doi.org/10.21609/jiki.v18i1.1440

[18] Prasetyo, B., & Al-Majid, A. Y. (2024). A Comparative Analysis of multinomialNB, SVM, and BERT on Garuda Indonesia Twitter sentiment. *Penelitian Ilmu Komputer Sistem Embedded and Logic*, *12*(2), 445–454. https://doi.org/10.33558/piksel.v12i2.9966

[19] Rifaldy, F., Sibaroni, Y., & Prasetiyowati, S. S. (2025). Effectiveness of Word2VEC and TF-IDF in sentiment classification on online investment platforms using support vector machine. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, *10*(2), 863–874. https://doi.org/10.29100/jipi.v10i2.6055

[20] Vincent, V., & Zahra, A. (2025). Multilingual hate speech detection using deep learning. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *14*(3), 1015. https://doi.org/10.11591/ijict.v14i3.pp1015-1023

[21] Ghozali, I., Sungkono, K. R., Sarno, R., & Abdullah, R. (2023). Synonym based feature expansion for Indonesian hate speech detection. *International Journal of Electrical and Computer Engineering*, *13*(1), 1105. https://doi.org/10.11591/ijece.v13i1.pp1105-1112

[22] Ahmad, M., Waqas, M., Hamza, A., Usman, S., Batyrshin, I., & Sidorov, G. (2025). UA-HSD-2025: Multilingual hate speech detection from tweets using pre-trained transformers. *Computers*, *14*(6), 239. https://doi.org/10.3390/computers14060239

[23] Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, *37*(1), 2166719. https://doi.org/10.1080/08839514.2023.2166719

[24] Kusuma, J. F., & Chowanda, A. (2023). Indonesian hate speech detection using IndoBERTweet and BiLSTM on Twitter. *International Journal on Informatics Visualization*, *7*(3), 773–780. https://dx.doi.org/10.30630/joiv.7.3.1035

[25] Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: A review. *Artificial Intelligence Review*, *56*(9), 10345–10425. https://doi.org/10.1007/s10462-023-10419-1

[26] Shreyas, S., Vathsal, V. S., & Shankar, S. R. (2024). Hate speech detection using deep learning models. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions*, 1–6. https://doi.org/10.1109/CSITSS64042.2024.10816708

[27] Hong-Phuc Vo, H., Nguyen, H. H., & Do, T. H. (2021). Online hate speech detection on Vietnamese social media texts in streaming data. In *International Conference on Artificial Intelligence and Big Data in Digital Era*, 315–325. https://doi.org/10.1007/978-3-030-97610-1_25

[28] Anitha, G., Karthik, G., Chadge, R., Dinesh, M., & Subha, T. D. (2024). A novel methodology design to predict hate speech on social media using machine learning principles. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems*, 1–6. https://doi.org/10.1109/ICSES63760.2024.10910624

[29] Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the third workshop on abusive language online*, (46-57). https://doi.org/10.18653/v1/W19-3506

[30] Brajesh Kumar, S., Dinesh Kumar, V., & Prateek, P. (2025). A novel framework for text preprocessing using NLP approaches and classification using random forest grid search technique for sentiment analysis. *Economic Computation*

and Economic Cybernetics Studies and Research, *59*(2/2025), 89–105. https://doi.org/10.24818/18423264/59.2.25.06

[31] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52. https://doi.org/10.3390/technologies9030052

[32] AbdelHamid, M., Jafar, A., & Rahal, Y. (2022). Levantine hate speech detection in Twitter. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/s13278-022-00950-4

[33] Salici, M., & Olcer, U. E. (2024). Impact of transformer-based models in NLP: An in-depth study on BERT and GPT. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1–6). https://doi.org/10.1109/idap64064.2024.10710796

[34] Shujaaddeen, A. A., Ba-Alwi, F. M., Zahary, A. T., & Alhegami, A. S. (2024). A model for measuring the effect of splitting data method on the efficiency of machine learning models: A comparative study. In *2024 4th International Conference on Emerging Smart Technologies and Applications*, 1–13. https://doi.org/10.1109/eSmarTA62850.2024.10639022

[35] Ibrahim, M. (2022). An explainable AI model for hate speech detection on Indonesian Twitter. *Communication and Information Technology Journal*, *16*(2), 175–182.

[36] Muhamad, S., Nico, G., & Permana, M. (2022). Indobert for Indonesian Fake News Detection. *ICIC Express Letters ICIC International*, *16*(3), 289–297. https://doi.org/10.24507/icicel.16.03.289

[37] Perifanos, K., & Goutsos, D. (2021). Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction*, *5*(7), 34. https://doi.org/10.3390/mti5070034

[38] Kamboj, P., Kumar, S., & Goyal, V. (2025). Reducing social biases in language models: A comparative evaluation of debiasing strategies. In *2025 2nd International Conference on Advanced Computing and Emerging Technologies*, 1–6. https://doi.org/10.1109/ACET67282.2025.11430255

How to Cite: Bastian, Y. I., Hermawan, A., Kurniawan Maranto, A. R., Daniawan, B., & Junaedi, J. (2026). Deep Learning Approaches for Hate Speech Detection in Resource-Constrained Social Media Texts: A Comparative Evaluation. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027379

AIA62027379_R1 (1).pdf

ORIGINALITY REPORT

12%
SIMILARITY INDEX

PRIMARY SOURCES

1	ojs.bonviewpress.com Internet	126 words — 2%
2	github.com Internet	50 words — 1%
3	Md. Khaliluzzaman, Kaushik Deb. "A Multi-Part Attention-Guided Spatial-Temporal GCN Framework for Gait-Based Person Recognition", Artificial Intelligence and Applications, 2026 Crossref	48 words — 1%
4	www.mdpi.com Internet	46 words — 1%
5	mdpi-res.com Internet	29 words — < 1%
6	cendikiainovasi.org Internet	27 words — < 1%
7	www.astesj.com Internet	22 words — < 1%
8	Aldi Pranadia, Rila Mandala. "Implementation of Query Expansion to Enhance Word2Vec Performance in Hoax News Detection Systems", 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), 2023 Crossref	21 words — < 1%
9	epa.oszk.hu Internet	20 words — < 1%
10	ejurnal.stmik-budidarma.ac.id Internet	19 words — < 1%

11	Fatih Fauzan Kartamanah, Aldy Rialdy Atmadja, Ichsan Budiman. "Analyzing PEGASUS Model Performance with ROUGE on Indonesian News Summarization", sinkron, 2025 Crossref	18 words — < 1%
12	V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Next-Generation Smart Systems - Bridging Sustainability and Computational Intelligence", CRC Press, 2026 Publications	18 words — < 1%
13	jurnal.atmaluhur.ac.id Internet	18 words — < 1%
14	"Proceedings of the 6th International Conference on Artificial Intelligence and Smart Energy", Springer Science and Business Media LLC, 2026 Crossref	17 words — < 1%
15	iieta.org Internet	16 words — < 1%
16	www.jisem-journal.com Internet	14 words — < 1%
17	ejurnal.stie-trianandra.ac.id Internet	13 words — < 1%
18	www.ojcmt.net Internet	13 words — < 1%
19	Meet Mehta, Dhruv Gada, Riddhi Sharma, Khushi Chavan, Pratik Kanani. "Offense Detection Using BERT and CNN", 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), 2022 Crossref	11 words — < 1%
20	Rameesha Zia, Muhammad Shahid Iqbal Malik. "Two-Level Violence Incitation Detection in Urdu Using Explainable Hybrid Deep Learning Model", Applied Soft Computing, 2026 Crossref	11 words — < 1%
21	Sarawut Ramjan, Jirapon Sunkpho. "chapter 5 Deep Learning", IGI Global, 2025 Crossref	11 words — < 1%
22	arxiv.org	

Internet

11 words — < 1%

23 "Digital Technologies and Applications", Springer Science and Business Media LLC, 2026
Crossref

10 words — < 1%

24 "Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track", Springer Science and Business Media LLC, 2021
Crossref

10 words — < 1%

25 Endang Wahyu Pamungkas, Divi Galih Prasetyo Putri, Azizah Fatmawati. "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities", International Journal of Advanced Computer Science and Applications, 2023
Crossref

10 words — < 1%

26 joiv.org
Internet

10 words — < 1%

27 "Proceedings of Data Analytics and Management", Springer Science and Business Media LLC, 2026
Crossref

9 words — < 1%

28 F Fayas Ahamed, M Prasanth, Atul Saju Sundaresh, D Manoj Krishna, S Sindhu. "Multimodal Hate Speech Detection With Explanability Using LIME", 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024
Crossref

9 words — < 1%

29 Jing Shen. "Corpus Classification Algorithm Based on BERT Pre-trained Model", Procedia Computer Science, 2025
Crossref

9 words — < 1%

30 Muhammad Amien Ibrahim, Noviyanti Tri Maretta Sagala, Samsul Arifin, Rinda Nariswari et al. "Separating Hate Speech from Abusive Language on Indonesian Twitter", 2022 International Conference on Data Science and Its Applications (ICoDSA), 2022
Crossref

9 words — < 1%

31 Muhammad Deedahwar Mazhar Qureshi, M. Atif Qureshi, Wael Rashwan. " Explainable for Hate Speech Moderation: A Stakeholder-Centered and Sociotechnical Review ", WIREs Data Mining and Knowledge Discovery, 2026
Crossref

32	essay.utwente.nl Internet	9 words — < 1%
33	link.springer.com Internet	9 words — < 1%
34	vtechworks.lib.vt.edu Internet	9 words — < 1%
35	"Proceedings of the Third International Conference on Advances in Computing Research (ACR'25)", Springer Science and Business Media LLC, 2025 Crossref	8 words — < 1%
36	Anis Charfi, Mabrouka Besghaier, Raghda Akasheh, Andria Atalla, Wajdi Zaghouani. "Hate speech detection with ADHAR: a multi-dialectal hate speech corpus in Arabic", Frontiers in Artificial Intelligence, 2024 Crossref	8 words — < 1%
37	Bhabani Prasanna Pattanaik, Dhirendra Kumar Shukla, Sambit Satpathy, Kamalakanta Muduli. "Sustainable Developments in Computer Engineering, Green Technology and Smart Systems", CRC Press, 2026 Publications	8 words — < 1%
38	Endang Wahyu Pamungkas, Arida Ferti Syafiandini, Dian Purworini, Widi Widayat et al. "Reading between modalities: multimodal hate speech detection in low-resource Indonesian social media", Journal of Computational Social Science, 2026 Crossref	8 words — < 1%
39	Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga. "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions", Computer Science Review, 2020 Crossref	8 words — < 1%
40	Igor Petrušić, Andrej Savić, Katarina Mitrović, Nebojša Bačanin, Gabriele Sebastianelli, Daniele Secci, Gianluca Coppola. "Machine learning classification meets migraine: recommendations for study evaluation", The Journal of Headache and Pain, 2024 Crossref	8 words — < 1%
41	N. V. R. Naidu, N. L. Ramesh, A. Jagannatha Reddy, Nagaraju Kottam et al. "Applied Research in Engineering Sciences", CRC Press, 2026 Publications	8 words — < 1%

42	Shynar Mussiraliyeva, Kalamkas Bagitova, Daniyar Sultan. "Social Media Mining to Detect Online Violent Extremism using Machine Learning Techniques", International Journal of Advanced Computer Science and Applications, 2023 Crossref	8 words — < 1%
43	ejournal.pnc.ac.id Internet	8 words — < 1%
44	library.iyte.edu.tr Internet	8 words — < 1%
45	library.poltekkes-surabaya.ac.id Internet	8 words — < 1%
46	www.nature.com Internet	8 words — < 1%
47	"Proceedings of International Conference on AI Systems and Sustainable Technologies", Springer Science and Business Media LLC, 2026 Crossref	7 words — < 1%
48	"Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning", Springer Science and Business Media LLC, 2024 Crossref	7 words — < 1%
49	Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, Amir Hussain. "Hate speech detection: A comprehensive review of recent works", Expert Systems, 2024 Crossref	7 words — < 1%
50	Shoffan Saifullah, Rafał Drężewski. "A Hybrid Feature-Enhanced IndoBERT Framework with Controlled Semi-Supervised Learning for Low-Resource Indonesian Hate Speech Detection", Applied Sciences, 2026 Crossref	7 words — < 1%
51	Guto Leoni Santos, Vitor Gaboardi dos Santos, Colm Kearns, Gary Sinclair et al. "All together against hate: ensemble based LLMs for multi class hate speech classification in the football context", Journal of Big Data, 2026 Crossref	6 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF