

Why simple neural networks outperform traditional classifiers in cyberbullying detection

Rini Anggrainingsih, Firdaus Ashiddqi Noor, Gita Bangun Prakoso, Wiharto, Dewi Wardani,
Hasan Dwi Cahyono

Department of Informatics, Faculty of Information Technology and Data Science, Sebelas Maret University, Surakarta, Indonesia

Article Info

Article history:

Received Nov 1, 2025

Revised Jul 25, 2026

Accepted Aug 12, 2026

Keywords:

Contextual word embeddings
Cyberbullying detection
Indonesian bidirectional encoder
representations from transformers
Neural vs. traditional classifiers
Text classification

ABSTRACT

Cyberbullying on social media, particularly in Bahasa Indonesia, has become a pressing problem. Detecting abusive language in Bahasa Indonesia poses unique challenges due to the prevalence of informal expressions, slang, code-mixing, and implicit aggression. Recently pre-trained language models, such as Indonesian bidirectional encoder representations from transformers (IndoBERT), can generate contextual word embeddings that have shown strong potential for understanding complex linguistic patterns. However, it remains unclear which classification models can best exploit these rich embeddings. This study aims to enhance the performance of cyberbullying detection and investigate why simple neural networks outperform traditional machine learning classifiers, such as support vector machine (SVM) and naïve Bayes (NB) when applied to contextual embeddings in Bahasa Indonesia. This study conduct both theoretical and empirical comparisons of fully connected neural networks (FCNN), SVM, and NB classifiers using IndoBERT embeddings. The results demonstrate that even shallow neural architectures can leverage contextual semantic information more effectively than traditional methods, yielding higher accuracy and robustness. Besides performance improvement, the study also explains the underlying reasons for this superiority, highlighting the adaptability of neural models to complex embedding spaces. These findings provide insights for developing more effective abusive language detection systems in low-resource languages and contribute to understanding the interaction between modern embeddings and classifier architectures.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rini Anggrainingsih

Department of Informatics, Faculty of Information Technology and Data Science, Sebelas Maret University

Ir Sutami St. 36A, Kentingan, Jebres, Surakarta 57126, Indonesia

Email: rini.anggrainingsih@staff.uns.ac.id

1. INTRODUCTION

Cyberbullying is an intentional, repeated, and harmful behavior enacted through digital channels, such as social media, messaging, games, and forums. It remains a serious concern on social platforms. Recent studies on cyberbullying reported that cyberbullying happens across many social media platforms and that it is still hard for computers to detect such behavior automatically [1].

The high cyberbullying rate in urban areas of Indonesia, such as Jakarta, Surabaya, and Bandung, underscores the urgent need for automatic detection systems tailored to Bahasa Indonesia [2]. On the other hand, the rapid growth of online discourse and the mixing of formal Bahasa Indonesia with slang and

code-mixing create additional challenges to build a highly accurate automatic detection for Bahasa Indonesia, reinforcing the practical need for language-aware models in moderation pipelines [3]. Recently, transformers contextual embeddings that can capture the meaning of words in context have become the standard tool for text classification tasks.

For Indonesian social networks, transformer-based encoders, such as Indonesian bidirectional encoder representations from transformers (IndoBERT), can be utilized to capture informal expressions and platform vernacular in Bahasa Indonesia more effectively than feature engineering methods, yielding consistent gains across text classification tasks. Recent Indonesian natural language processing (NLP) studies on Twitter (now X) further underscore the benefit of transfer learning and contextual encoders for downstream classification quality. These trends motivate the use of Indonesian contextual representations as features for cyberbullying detection to classify a harmful or abusive text in online conversations [4], [5].

While transformer can represent text into contextual embeddings that supply powerful features, the choice of classifier over those features remains essential for gaining high performance and efficiency. Comparative benchmarks in some studies show that linear support vector machines (SVMs) are strong and often competitive baselines for text and naïve Bayes (NB) remains attractive for speed and low-resource settings in the cyberbullying field in Bahasa Indonesia [6], [7]. In the context of cyberbullying detection in Bahasa Indonesia, most prior works more emphasized improving detection accuracy rather than understanding why deep learning models, such as simple neural networks, better utilize contextual embeddings than traditional classifiers. To address this gap, the present study investigates how neural networks capture contextual information and interact with different classifier architectures by comparing the performance of SVM, NB, and fully connected neural layers built on top of IndoBERT in classifying cyberbullying text in Bahasa Indonesia.

Specifically, in this study: i) employs IndoBERT as a feature extractor to generate Indonesian text representations for cyberbullying detection; ii) evaluates and compares the performance of SVM, NB, and fully connected neural layers built on top of IndoBERT in classifying cyberbullying text in Bahasa Indonesia; and iii) analysis, from both theoretical and practical perspectives on how the choice of classifier impacts overall classification performance. Through these analysis, this study aims to provide a deeper understanding for researchers and practitioners in selecting lightweight yet high-precision architectures. This study also contributes to the development of more robust multilingual safety systems for online text moderation in Bahasa Indonesia.

2. LITERATURE REVIEW

2.1. Characteristics of cyberbullying in Bahasa Indonesia

In the Bahasa Indonesia context, cyberbullying has different unique features compared to English or other monolingual settings. In English language, offensive words often follow more standardized patterns, while social media posts in Bahasa Indonesia are dynamic, informal, creative, and often mix with other languages. People often use mixed languages, slang, sarcasm, or subtle hints to express harmful messages without using explicit bad words. This makes it much harder for computer models to detect abusive language accurately.

One of the most notable features of Bahasa Indonesia in online communication is the use of code-mixing, a combination of Bahasa Indonesia and other languages, either content of Bahasa Indonesia-English or Bahasa Indonesia-local languages such as Javanese or Sundanese. Some studies reported that this multilingual and code-mixed usage is normal and often used in many Indonesian social networks to emphasize tone, create a sense of belonging to the group, or make sarcasm more subtle [8]. For example, a comment such as “*Dasar kamu toxic banget sih, literally bikin orang stress*” mixes Bahasa Indonesia with English terms. Another variation might include Javanese phrases, creating a trilingual mix within a single sentence. Indonesian youths frequently use informal slang and spellings on platforms like TikTok, Instagram, or X. For example:

- ‘*anjir*’ or ‘*anjay*’ as softened/slang forms of bad language for ‘*anjing*’ (dog, often used as an insult).
- ‘*baper*’ (from *bawa perasaan*, meaning overly emotional).
- ‘*gabut*’ (having nothing to do, often used negatively in teasing contexts).

In addition, spellings are often manipulated or twisted to elude moderation or make the bullying sound playful: ‘*tolol*’ instead of ‘*tolol*’ (stupid), or elongated words like ‘*sumppaaaahhh*’ for emphasis. These kinds of linguistic ways make dictionary-based detection not reliable anymore, since new variants emerge continuously [5].

Furthermore, there is often a sarcasm, another type of cyberbullying in Bahasa Indonesia. A user may write “*Wah pinter banget ya, nilainya merah semua*” (“Wow, you are so smart, all your grades are failing”). While no explicit insult is present, the sarcastic context makes it harmful. Sarcasm often leverages positive surface words to mask negative intent, making keyword-based approaches no longer effective. Recent studies confirmed that sarcasm detection benefits from contextual and sentiment-aware models [9]. However, in Bahasa Indonesia, a sarcasm case is challenging because sarcasm often relies on cultural references, such as local humor, memes, or trending hashtags, that are not easily transferable from English-language models.

2.2. How contextual embeddings like IndoBERT capture nuanced meaning and why traditional features fall short

Conventional text representations and feature engineering treat words as isolated units, ignoring context and order. For example, term frequency-inverse document frequency (TF-IDF), word to vectors or bag-of-words. Despite being computationally cheap, traditional text representations are no longer adequate for abusive-language detection in Indonesian social media due to the following reasons [5], [10]:

- Context loss, as a traditional text representation cannot differentiate whether a praise sentence is sarcastic compliments or genuine.
- Sparsity issues due to a continuing emerging informal vocabulary, such as slang, spelling variants, and creative expressions vocabulary, a traditional representation method failed to recognise a new slang vocabulary early.
- Code-mixing failures as words from different languages are treated as unrelated tokens, reducing the ability to capture cross-lingual bullying cues.

In contrast, contextual embeddings produced by transformer-based architectures encode meaning dynamically by considering both directions of the surrounding words, the left and right of a word. This ability is valuable, particularly for cyberbullying detection in Bahasa Indonesia. For example, in the sentence “*Dia memang gila belajar*” (“She’s crazy about studying”), the word “*gila*” (crazy) is positive, denoting enthusiasm. In “*Dia gila bener, bikin malu keluarga*” (“He’s really crazy, bringing shame to the family”), it is a negative insult. A TF-IDF model would treat “*gila*” as the same token, but transformer models, such as IndoBERT can distinguish its meaning based on context by considering words before and after “*gila*” [5], [11]. IndoBERT is a pre-trained language model on a large Indonesian corpus, IndoBERT captures local slang and idiomatic usage better than multilingual models. Studies report significant accuracy improvements in emotion and sentiment classification tasks compared to baseline models [4].

2.3. Theoretical comparison of classifiers

The high performance and reliability of text classification in Bahasa Indonesia depend not only on the text representation method but also on the choice of classifier model. This section presents a comparative analysis of three classifiers: fully connected neural networks / multi-layer perceptron (FCNN / MLP), SVM, and NB. These three selected classifiers represent distinct learning paradigms: probabilistic (NB), margin-based (SVM), and representation-based (FCNN), allowing a balanced comparison between traditional and neural approaches in leveraging contextual embeddings for cyberbullying detection. The comparison covers both theoretical foundations and practical trade-offs.

2.3.1. Fully connected neural networks / multi-layer perceptron

An FCNN (often called MLP) is the simplest type of artificial neural networks where all neurons in one layer are connected to every neuron in both the previous and the next layers. This composition of neurons enables the model to learn a mapping function $y = f(x)$ between input x and output y through a series of linear transformations, each followed by a nonlinear activation function [12], [13]. For a classification case, FCNN takes a feature vector and passes it through layers of linear transforms and nonlinear activations, and a softmax for predicting a class at the end.

When the input is a contextual embedding (for example, mean-pooled token states from IndoBERT), the FCNN can learn how to mix those high-level features. It will be able to boost or suppress dimensions and combine them multiplicatively through nonlinearities, and create interaction terms that approximate complex rules [14]. In text classification tasks, FCNNs are typically trained to minimize a loss function, often incorporating label smoothing or class weighting to handle class imbalance. To achieve optimal performance, the network can be regularized and optimized through techniques such as dropout, weight decay, early stopping, and batch normalization.

2.3.2. Naïve Bayes

The NB classifier is a probabilistic supervised learning algorithm that is built based on the foundation of Bayes' Theorem, which states as in (1).

$$P(Y = c | X = x) = \frac{P(Y = c) P(X = x | Y = c)}{P(X = x)} \quad (1)$$

In the classification setting, Y refers to the class label and X refers to the vector of random variables that represent the features. The "naive" assumption is that the features are conditionally independent given the class label, as in (2).

$$P(C | X) \propto P(C) \prod_i P(x_i | C) \quad (2)$$

This assumption can reduce the number of parameters to estimate and simplify computation, making the classifier much more efficient [15]. NB has been reported to gain high performance in many real-world applications, especially when feature dimensionality is high. This includes text classification, spam detection, and other sparse-high-dimensional domains where dependencies have a limited negative impact [16].

2.3.3. Support vector machines

SVM is one of the supervised-learning algorithms that is inspired by statistical learning theory, hence benefiting SVM to have strong generalization performance when the margin is large [17]. The basic idea of SVM is to find the best decision border (hyperplane) that separates classes of data by maximizing margin from the hyperplane to the nearest training points of any class [18], [19]. SVM control capacity by considering the distance from the border to the nearest point rather than the number of parameters. Therefore, SVM would be effective in high-dimensional spaces or when the number of dimensions is large relative to the number of samples [17].

In addition, because margin maximization implicitly imposes capacity control, SVMs often generalize well, particularly when properly regularized; hence, SVM is robust against over-fitting. However, although SVM establish a clear border (hyperplane), the impact of individual features is often less considerable compared to simpler models [20]. Moreover, margin-based method in SVM may struggle and need extensive parameter tuning when classes are highly overlapping or contain many outliers [21].

2.4. Cyberbullying detection in Bahasa Indonesia

Previous studies on cyberbullying detection in Bahasa Indonesia have primarily employed conventional machine learning techniques to determine whether a text contains bullying elements. These methods typically classify text by extracting linguistic or statistical features and representing the text using approaches such as TF-IDF and bag-of-words, as seen in works by [6], [7]. More recent research leveraged emerging bidirectional encoder representations from transformers (BERT)-based models as feature extractors to encode text into contextual embeddings, as shown in studies such as [22], [23]. This advancement is allowing next other studies to compare classifier performance in identifying bullying content [24], [25].

In the context of Bahasa Indonesia, several studies have been published on cyberbullying detection. However, only a few have made their datasets publicly available. Among these, Lukman and Novianto [6] utilized a new dataset specifically designed for cyberbullying detection in Bahasa Indonesia publicly available. Despite these advances, most prior works focused on improving detection accuracy without fully explore why certain models—particularly deep learning architectures, including a simple neural networks tend to leverage contextual embeddings more effectively than traditional classifiers.

This study aims to address this gap by investigating the underlying factors that enable neural networks to capture contextual information during the embedding representation process and how they interact with different classifier architectures. Specifically, it seeks to explain why simple neural networks can leverage embeddings effectively and achieve better performance in Indonesian cyberbullying detection tasks. This study use a dataset similar to that employed in the work of Lukman and Novianto [6] and adopt it as the baseline for this study.

3. MATERIAL AND METHOD

3.1. Dataset

This study utilized a dataset of Indonesian-language social media comments that were annotated whether a text contained cyberbullying context or not. The dataset is released by an independent contributor on public data repository [26]. Specifically, the dataset used in this study contained 650 comments in total, which consisted of 325 instances of cyberbullying labeled and 325 non-bullying labeled samples. Although this study can take advantage of the balanced nature of this dataset to support fairness training and evaluation, this dataset is considered a relatively small size, also presents challenges. Small datasets in low-resource languages like Bahasa Indonesia may restrict the generalizability of models compared to large-scale benchmarks in English [3]. Hence, our empirical design focuses on maximizing the representational strength of embeddings (IndoBERT) while carefully comparing classifiers under identical protocols.

3.2. Preprocessing and embedding strategy

In this study conducted a data preprocessing strategy to remove noise and to standardize text before the encoding process. The text preprocessing steps include:

- Lowercasing: all text was converted into lowercase to get similar representations of similar words (e.g., “Cyberbullying” vs. “cyberbullying”).
- Removing mentions and hashtags as social media texts often include @username mentions and #hashtags that may carry little semantic value for bullying detection.
- Removing or normalizing Non-American Standard Code for Information Interchange (ASCII), extra symbols, emojis, and foreign characters. Emojis were converted to textual descriptors where possible, as they sometimes convey affective signals relevant to bullying.
- Normalizing token by applying a standard NLP cleaning operations, including removal of punctuation, stop-word filtering, and white-space normalization.

The processed texts were then tokenised and encoded into vector representations using IndoBERT, a pre-trained language model specifically trained on Indonesian corpora [27]. IndoBERT is essentially a BERT [28] that is trained on a large corpus of Indonesian news, Wikipedia, and web data, enabling it to capture semantic nuances in Indonesian texts that traditional embeddings, such as TF-IDF or Word2Vec, often fail to represent. Similarly to BERT, in IndoBERT, for each input sentence, the [CLS] token embedding was extracted from IndoBERT’s final hidden layer. This vector is then used as a contextual text representation of all sentences in the dataset to train the classifier models.

3.3. Classifier configurations and training setup

Previous studies have demonstrated that contextual embeddings can produce highly separable representations in which relatively simple classifiers can achieve strong text classification performance, supporting the use of linear SVM and Gaussian naïve Bayes (GNB) as a competitive baseline in embedding-based classification tasks [29], [30]. In this study compared three classifiers: a Feedforward neural network (FNN), a SVM, and a GNB. The FNN was implemented by leveraging IndoBERT’s embedded classification layer by feeding its contextual embeddings directly into the built-in classifier on the top layer of IndoBERT.

This study set random seed to 42, used adaptive moment estimation (Adam) as the optimizer with a learning rate of 2×10^{-5} and a batch size of 16, optimizing categorical cross-entropy for up to 50 epochs with early stopping based on the validation F_1 score to prevent over-fitting. This setup minimized additional architectural assumptions while allowing the model to exploit the representational capacity of IndoBERT embeddings for classification. For SVM, this study employed a linear kernel with the regularization constant C that was optimized during the training process.

This study implemented all classifier models in Python by using the Scikit-learn libraries. For splitting data, this study applied an 80 : 20 train-test split rather than a three-way split (for example, 70 : 20 : 10 for train, validation, and test). This study note that a separate validation set was not required because model selection was handled within the training process. In particular, FNN used early stopping based on the validation F1-score drawn from the training partition, while SVM and NB were trained directly using the same split without additional validation. This setup enabled us to allocate more data for training (80%), which is especially important in the low-resource context of Bahasa Indonesia data, while still having a sufficient number of test set (20%) to obtain a reliable estimate of generalisation performance.

At the final step, this study assessed the performance of each model using four standard evaluation metrics: accuracy, precision, recall, and F1-score using (3) to (6). These metrics provide a comprehensive view of classification quality by jointly reflecting correctness, sensitivity, and balance between precision and recall. In this context, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives used in the computation of these measures, respectively. This study run all experiments three times and record the best result.

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision(P) = \frac{TP}{TP + FP} \quad (4)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2(P)(R)}{P + R} \quad (6)$$

4. RESULTS AND DISCUSSION

This section presents our experimental results and analysis their implications for model behavior and representation learning. This study observe consistent gains when using embedding-based features, with improvements spanning accuracy, precision, recall, and F1-score across multiple splits and classifiers as presented in Tables 1 and 2. Building on these findings, this study develop a focused discussion on why deep learning models, including shallow FNN, leverage embeddings more effectively than traditional baselines.

Table 1. The performance of baseline model; cyberbullying detection using TF-IDF and bag-of-words

Models	Accuracy	Precision	Recall	F1-score
NB	0.82	0.75	0.88	0.81
SVM	0.88	0.88	0.88	0.88

Table 2. The performance of cyberbullying detection classifier models using contextual embedding

Models	Accuracy	Precision	Recall	F1-score
NB	0.777	0.704	0.919	0.798
SVM	0.831	0.813	0.839	0.826
IndoBERT	0.859	0.842	0.874	0.858
Finetuned IndoBERT	0.946	0.934	0.972	0.958

4.1. Experimental results

This study confirm the expected performance hierarchy across classifier families as mentioned in theory and previous studies in the text classification area [31], [32]. NB under-performs due to its independence assumptions being misaligned with dense, correlated sentence embeddings, yielding lower precision despite relatively high recall. SVM improves on NB by maximizing margins in high-dimensional spaces, but remains limited by the non-task-adapted geometry of frozen embeddings.

In contrast, a linear head top of IndoBERT already surpasses classical learners, reflecting the strong priors encoded by contextual representations. The best results are obtained with fine-tuning of IndoBERT, which reshapes the representation manifold to enhance separability between classes, thereby improving all metrics. The results demonstrate that by fine-tuning IndoBERT, our model outperforms the baseline model by improving accuracy by 6.6%. These findings indicate that fine-tuned BERT-family models reliably outperform SVM and NB for supervised text classification, including in Indonesian-language settings.

This study did not employ NB and SVM on fine-tuned IndoBERT embeddings because the strength of contextual embeddings in pretrained language models comes from end-to-end fine-tuning, which jointly optimizes the encoder and a simple discriminative network under the cross-entropy objective to reshape the embedding space so that classes become more separable for the downstream task. For this reason and based on the previous works report [33], this study have the opinion that substituting a post-hoc NB/SVM cannot

further adapt the encoder and typically adds only an alternative boundary over features already optimized for linear classification. Hence, adding NB/SVM on top of already fine-tuned IndoBERT embeddings is methodologically redundant and unlikely to yield additional scientific insight or measurable gains beyond the standard fine-tuned head [34].

4.2. Why deep learning (including shallow FNNs) leverages embeddings more effectively

In text classification, pretrained language models (e.g. BERT models) provide dense embeddings that already capture rich context and semantics. Then, during the fine-tuning process, BERT models can reshape the embedding space through shallow neural adaptation to generate a better text representation quality. Hence, neural networks, even when the classifier is only a simple network, can modify the internal composition of embeddings to enhance decision boundaries and inter-class separability, while traditional machine learning, such as SVM and NB, relies on static representations and manual feature engineering. This notion is in line with previous studies in their comprehensive survey of text classification approaches [35], [36].

From a theoretical perspective, several studies have formally proved that deep learning models benefit from the implicit bias of gradient descent. When this optimization is applied to logistic or cross-entropy loss on separable data, the iterates tend to converge toward the maximum-margin separator. Consequently, even a simple linear network can naturally seek large-margin, well-generalizing solutions once the embedding space is properly structured [37], [38].

Another important factor is the capacity of deep learning models for nonlinear feature mixing. A shallow FNN built on top of IndoBERT can re-weight the embedding dimensions to correct residual distortions and effectively learn a task-specific transformation of the feature space. In contrast, traditional classifiers operate on static features and cannot modify the underlying geometry [39]. These findings conclude that the surpassing performance of Transformer models, such as IndoBERT, is driven by the fusion between fine-tuned embeddings and adaptive neural heads, not by the classifier's form.

5. CONCLUSION

This study demonstrates that simple neural-based classifiers consistently achieve higher performance across all evaluation metrics compared to traditional machine learning models on contextual embedding vectors. Specifically, fine-tuned IndoBERT achieved the highest F1-score (0.958), outperforming SVM (0.826) and NB (0.798). This study's findings confirmed the findings from the previous studies that the strength of contextual embeddings is gained most through end-to-end fine-tuning and adaptive neural layers, which can reshape the embedding space. The fine-tuning process enables neural layers to emphasize inter-class separability and reduce representational distortion. SVM and NB, operating on static feature representations, so that they do not have the ability to exploit the nonlinear interactions and semantic richness that may exist in the embedding vectors. Overall, this article contributes to the research on cyberbullying detection by highlighting the compatibility between contextual embeddings and neural classifiers. This study can be extended by future research by exploring multilingual and multimodal embeddings, larger and more diverse datasets, to enhance transparency and robustness in cyberbullying detection systems.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Sebelas Maret University and Data Information and Knowledge Engineering (DIKE) Research Group and for their outstanding support and sponsorship of this research.

FUNDING INFORMATION

This study was funded by Sebelas Maret University through the Research Group Grant under contract number 371/UN27.22/PT.01.03/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rini Anggrainingsih	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Firdaus Ashiddiqi Noor		✓	✓			✓		✓	✓					
Gita Bangun Prakosa		✓	✓					✓	✓		✓			
Wiharto		✓		✓						✓		✓		
Dewi Wardani	✓									✓			✓	✓
Hasan Dwi Cahyono	✓									✓			✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. There are no financial or personal relationships that could inappropriately influence or bias the content of this work.

INFORMED CONSENT

Not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

Not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely on computational models, publicly available datasets, or non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or animal ethics committee was not necessary for this study.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

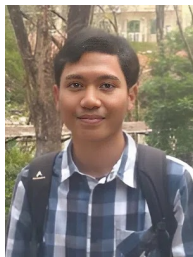
REFERENCES

- [1] V. Balakrisnan and M. Kaity, "Cyberbullying detection and machine learning: a systematic literature review," *Artificial Intelligence Review*, vol. 56, no. S1, pp. 1375–1416, Oct. 2023, doi: 10.1007/s10462-023-10553-w.
- [2] E. L. Amanatin and B. Sekarningrum, "Causes and forms of cyberbullying among teenagers in Indonesian urban areas: cases of Jakarta, Bandung and Surabaya," *JISPO Jurnal Ilmu Sosial dan Ilmu Politik*, vol. 13, no. 2, pp. 233–254, Mar. 2024, doi: 10.15575/jispo.v13i2.27792.
- [3] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: review of the state of the art," *Information Processing & Management*, vol. 60, no. 5, Sep. 2023, doi: 10.1016/j.ipm.2023.103454.
- [4] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT," *Social Network Analysis and Mining*, vol. 15, no. 1, Mar. 2025, doi: 10.1007/s13278-025-01439-6.
- [5] M. O. Ibrahim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: progress and challenges," *Heliyon*, vol. 9, no. 8, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18647.
- [6] K. Lukman and S. Novianto, "Komparasi algoritma naïve Bayes dan SVM untuk identifikasi cyberbullying selebriti di media sosial Twitter," *Jurnal Algoritma*, vol. 22, no. 1, pp. 970–981, May 2025, doi: 10.33364/algoritma/v.22-1.2196.
- [7] A. Alabdulwahab, M. A. Haq, and M. Alshehri, "Cyberbullying detection using machine learning and deep learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 424–432, 2023, doi: 10.14569/IJACSA.2023.0141045.
- [8] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Word level language identification in Indonesian-Javanese-English code-mixed text," *Procedia Computer Science*, vol. 244, pp. 105–112, 2024, doi: 10.1016/j.procs.2024.10.183.
- [9] O. Vitman, Y. Kostiuk, G. Sidorov, and A. Gelbukh, "Sarcasm detection framework using context, emotion and sentiment features," *Expert Systems with Applications*, vol. 234, Dec. 2023, doi: 10.1016/j.eswa.2023.121068.
- [10] M. S. Jahan and M. Ouassalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, doi: 10.1016/j.neucom.2023.126232.

- [11] Anjum and R. Katarya, "Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities," *International Journal of Information Security*, vol. 23, no. 1, pp. 577–608, Feb. 2024, doi: 10.1007/s10207-023-00755-2.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning (adaptive computation and machine learning series)*. Cambridge, US: The MIT Press, 2016.
- [13] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, Mar. 2021, doi: 10.1145/3446776.
- [14] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.
- [15] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.
- [16] G. Feng, J. Guo, B.-Y. Jing, and T. Sun, "Feature subset selection using naive Bayes for text classification," *Pattern Recognition Letters*, vol. 65, pp. 109–115, Nov. 2015, doi: 10.1016/j.patrec.2015.07.028.
- [17] J. Cervantes, F. G.-Lamont, L. R.-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.
- [18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [19] A. Shmilovici, "Support vector machines," in *Machine Learning for Data Science Handbook*, Cham, Switzerland: Springer International Publishing, 2023, pp. 93–110, doi: 10.1007/978-3-031-24628-9_6.
- [20] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring kernel machines and support vector machines: principles, techniques, and future directions," *Mathematics*, vol. 12, no. 24, Dec. 2024, doi: 10.3390/math12243935.
- [21] K. Shyrokykh, M. Giryk, and L. Dellmuth, "Short text classification with machine learning in the social sciences: the case of climate change on Twitter," *PLOS ONE*, vol. 18, no. 9, Sep. 2023, doi: 10.1371/journal.pone.0290762.
- [22] M. A. S. Nasution and E. B. Setiawan, "Enhancing cyberbullying detection on Indonesian Twitter: leveraging FastText for feature expansion and hybrid approach applying CNN and BiLSTM," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 929–936, Aug. 2023, doi: 10.18280/ria.370413.
- [23] Al-Khowarizmi, I. P. Sari, and H. Maulana, "Detecting cyberbullying on social media using support vector machine: a case study on Twitter," *International Journal of Safety and Security Engineering*, vol. 13, no. 4, pp. 709–714, Sep. 2023, doi: 10.18280/ijss.130413.
- [24] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in *2020 International Conference on Electronics and Sustainable Communication Systems*, Jul. 2020, pp. 1096–1100, doi: 10.1109/ICESC48915.2020.9155700.
- [25] H. Saini, H. Mehra, R. Rani, G. Jaiswal, A. Sharma, and A. Dev, "Enhancing cyberbullying detection: a comparative study of ensemble CNN-SVM and BERT models," *Social Network Analysis and Mining*, vol. 14, no. 1, Dec. 2023, doi: 10.1007/s13278-023-01158-w.
- [26] C. T. Hanni, "Cyberbullying Bahasa Indonesia," kaggle.com, 2023. Accessed Date: Dec. 23, 2025. [Online]. Available: <https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia/data>
- [27] B. Wilie et al., "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
- [29] C. M. V. D. Andrade, F. M. Belém, W. Cunha, C. França, F. Viegas, L. Rocha, and M. A. Gonçalves, "On the class separability of contextual embeddings representations," *Information Processing & Management*, vol. 60, no. 4, 2023, doi: 10.1016/j.ipm.2023.103336.
- [30] S. S.-U.-Sakib, M. R. Rahman, M. S. A. Forhad, and M. A. Aziz, "Cyberbullying detection of resource constrained language from social media using transformer-based approach," *Natural Language Processing Journal*, vol. 9, Dec. 2024, doi: 10.1016/j.nlp.2024.100104.
- [31] S. Talebi, E. Tong, A. Li, G. Yamin, G. Zaharchuk, and M. R. K. Mofrad, "Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12911-024-02444-z.
- [32] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *Journal of Big Data*, vol. 10, no. 1, Jun. 2023, doi: 10.1186/s40537-023-00781-w.
- [33] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," in *Chinese Computational Linguistics*, Cham, Switzerland: Springer, 2019, pp. 194–206, doi: 10.1007/978-3-030-32381-3_16.
- [34] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: what we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, Dec. 2020, doi: 10.1162/tacL.a.00349.
- [35] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, Apr. 2021, doi: 10.1016/j.future.2020.11.022.
- [36] Q. Li et al., "A survey on text classification: from traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022, doi: 10.1145/3495162.
- [37] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.
- [38] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry, "Convergence of gradient descent on separable data," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019, pp. 3420–3428.
- [39] Y.-G. Xu, X.-P. Qiu, L.-G. Zhou, and X.-J. Huang, "Improving BERT fine-tuning via self-ensemble and self-distillation," *Journal of Computer Science and Technology*, vol. 38, no. 4, pp. 853–866, Jul. 2023, doi: 10.1007/s11390-021-1119-0.

BIOGRAPHIES OF AUTHORS

Rini Anggrainingsih    earned her Ph.D. from the University of Western Australia. She currently serves as a researcher and lecturer at the Faculty of Information Technology and Data Science, Sebelas Maret University. Her current research interests are focused on information systems, business process management, software quality, and data quality. The goal is to develop and optimize information systems to enhance organizational efficiency, analyze and improve business processes to align with strategic objectives, ensure software reliability and maintainability, and improve data accuracy and integrity. Emphasis is placed on integrating these areas to support data-driven decision-making and deliver robust solutions to contemporary challenges, particularly in combating the spread of false information and cyberbullying. She can be contacted at email: rini.anggrainingsih@staff.uns.ac.id.



Firdaus Ashiddiqi Noor    is a recent Informatics graduate from Sebelas Maret University, Indonesia, with an interest in artificial intelligence, particularly in NLP. He is involved project exploring detecting cyberbullying using Indonesian BERT. He can be contacted at email: firdashiddiqin@student.uns.ac.id.



Gita Bangun Prakoso    is a recent Informatics graduate from Sebelas Maret University, Indonesia, with a strong interest in artificial intelligence, particularly in NLP. Throughout his academic journey, he explored diverse fields, from web and mobile development to user experience design, before finding his passion in artificial intelligence. His undergraduate manuscript focused on detecting online gambling promotion comments on YouTube using Indonesian BERT and Multilingual BERT, involving large-scale data collection, manual annotation, and model fine-tuning for classification. He can be contacted at email: gpratama477@student.uns.ac.id.



Wiharto    is a professor in the Department of Informatics, Faculty of Information Technology and Data Science, Sebelas Maret University, Surakarta. He has a strong academic background and extensive experience in the field of computational science and engineering. He earned his doctoral degree from Universitas Gadjah Mada, where he specialized in biomedical informatics. His research primarily focuses on biomedical informatics, artificial intelligence, and computational intelligence, with an emphasis on solving real-world problems through innovative computational approaches. His contributions extend to mentoring students, publishing impactful research, and collaborating on interdisciplinary projects to advance the fields of health informatics and artificial intelligence. He can be contacted at email: wiharto@staff.uns.ac.id.



Dewi Wardani    earned her doctoral degree from Johannes Kepler University and completed her master's studies at National Cheng Kung University in Taiwan. Currently, she serves as a lecturer and researcher at the Faculty of Information Technology and Data Science, Sebelas Maret University in Surakarta, Central Java, Indonesia. Her research focuses on information systems, particularly knowledge management using a semantic web approach. Her work spans all layers of this field, from foundational concepts to intermediate and advanced applications. Additionally, she has a strong interest in language computation, which plays a vital role in addressing semantic challenges in knowledge management. She can be contacted at email: dww_ok@uns.ac.id.



Hasan Dwi Cahyono    is a lecturer at the Faculty of Information Technology and a member of the Data Science group at Sebelas Maret University in Surakarta, Central Java, Indonesia. His research interests focus on software engineering, generative modeling, and normalizing flows, which play a key role in enabling distribution transformations within generative models. He can be contacted at email: hasan.dwi.cahyono@gmail.com.