Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1127

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Analysis of Provocative Speech During the 2025 DPR Demonstration on X Using the IndoBERTweet Method

Nazhrin Nazarudin Achmad[1*]✉, Yuliant Sibaroni[1], Sri Suryani Prasetyowati[1]
[1]Faculty of Informatics, Universitas Telkom, Bandung, Indonesia
Corresponding Author: authors@ft.unp.ac.id

Article Information

Article history:

No. 1127
Rec. March 17, 2026
Rev. May 11, 2026
Acc. May 19, 2026
Pub. May 25, 2026
Page. 1600 – 1615

Keywords:

- *Provocative Speech*
- *IndoBERTweet*
- *Text Classification*
- *Natural Language Processing*
- *Social Media Analysis*

ABSTRACT

Social media platforms have become important channels for public discussion during political events. During the DPR demonstrations in August 2025, online discussions on X (formerly Twitter) contained various forms of expressions, including provocative speech that may influence public opinion and collective behavior. Detecting such content automatically is challenging due to the informal language, slang, and contextual nuances commonly found in social media texts. This study aims to analyze provocative speech on the social media platform X using text classification techniques and transformer-based models. A total of 8,899 Indonesian tweets related to the demonstration period from August 25 to August 31, 2025 were collected using the Tweet Harvest crawling tool. The dataset was manually labeled into two categories, namely provocative and non-provocative, using a majority voting approach by three annotators. Several preprocessing steps were applied, including cleaning, normalization, stemming, tokenization, and stopword removal. Several models were evaluated, including Multinomial Naïve Bayes, Linear Support Vector Machine, BiLSTM, IndoBERT, and IndoBERTweet. Experimental results show that transformer-based models outperform traditional machine learning approaches. The best performance was achieved by the IndoBERTweet model with a learning rate of 3×10^{-5}, achieving an accuracy of 93.07% and an F1-score of 91.56%. These findings indicate that domain-specific language models are effective for detecting provocative speech in Indonesian social media discussions related to political events.

How to Cite:

Achmad, N. N., & et al. (2026). Analysis of Provocative Speech During the 2025 DPR Demonstration on X Using the IndoBERTweet Method. Jurnal Teknologi Informasi Dan Pendidikan, 19(2), 1600-1615. https://doi.org/10.24036/jtip.v19i2.1127

Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1127

1. INTRODUCTION

The rapid growth of social media has transformed the way individuals express opinions and shape public discourse. Platforms like X (formerly Twitter) allow users to distribute information instantly and interact with large audiences in real time, making them influential spaces for socio-political discussions. During major social or political events, online discussions often intensify and may contain emotionally charged or provocative expressions that can influence public perception and collective behavior. For instance, during the DPR demonstrations in August 2025, social media platforms became a major channel for public expression, where narratives related to protest escalation and collective mobilization rapidly circulated online [1], [2].

In such contexts, provocative speech may emerge and potentially influence public reactions and social dynamics. Provocative speech refers to expressions intended to incite or encourage particular actions, often in ways that may escalate tension or conflict [3]. Unlike ordinary criticism or general hate speech, provocative content may contain elements of mobilization, persuasion, or incitement toward collective action. While hate speech generally focuses on attacking or discriminating against individuals or groups, provocative speech emphasizes stimulating emotional reactions or encouraging collective behavior within certain socio-political contexts. In contrast, sentiment analysis primarily aims to identify emotional polarity, such as positive, negative, or neutral opinions, rather than detecting persuasive or conflict-escalating intentions. Computationally, provocative speech detection requires models to capture contextual intent and implicit linguistic patterns beyond simple emotional classification. Identifying such content is challenging due to the characteristics of social media text, which frequently contain informal language, slang, abbreviations, and contextual nuances that are difficult for automated systems to interpret accurately.

Recent advancements in Natural Language Processing (NLP) enable the automatic analysis of substantial quantities of textual data from social media, particularly with the adoption of transformer-based models that have demonstrated superior performance in various classification tasks [4]. Text classification techniques are frequently employed to categorise user-generated content into established classifications. These methodologies have been effectively employed in numerous studies, encompassing sentiment analysis, opinion mining, and hate speech detection within social media data. Machine learning and

deep learning methodologies have been extensively utilised to examine extensive textual data from social media platforms.

Previous studies on Indonesian text classification have primarily focused on sentiment analysis and hate speech detection, with several works highlighting both the progress and remaining challenges in this area [5], [6]. For instance, Kusuma and Chowanda [7] investigated hate speech detection on the social media platform X using a combination of IndoBERTweet and BiLSTM, achieving an accuracy of 93.7%. Similarly, Dwitama et al. [8] conducted a study on Indonesian hate speech detection by optimizing a BiLSTM model through hyperparameter tuning and class weighting techniques, resulting in an accuracy of 97.66%. Despite these promising results, both studies still faced challenges related to class imbalance in certain categories and the potential risk of overfitting.

Other studies have also explored sentiment analysis on social media using transformer-based models. Dewi et al. [9] and Sayarizki et al. [10] examined public sentiment regarding social and political issues, such as the boycott movement against pro-Israel products and public opinion toward presidential candidates. These studies employed IndoBERT along with lexicon-based labeling and fine-tuning approaches. Their findings demonstrated that transformer-based models such as IndoBERT are highly effective for Indonesian sentiment classification and can achieve strong evaluation performance across various social media datasets.

However, research specifically focusing on provocative speech remains relatively limited, particularly in the context of Indonesian social media discussions related to political events and demonstrations. Most existing studies primarily concentrate on hate speech detection or sentiment analysis, while provocative expressions that may encourage conflict or mobilization have not been extensively examined. Therefore, this study aims to analyze provocative speech on the social media platform X using text classification techniques and transformer-based models. By evaluating traditional machine learning models and the IndoBERTweet model, this research seeks to better understand patterns of provocative discourse during demonstration-related discussions on social media.

2. RESEARCH METHOD

This study adopts an experimental approach to analyze and detect provocative speech on the social media platform X using several text classification models. The overall research process consists of multiple stages, including data acquisition, data preparation, model development, and evaluation. These stages aim to systematically process social media data and evaluate the effectiveness of different classification approaches for detecting provocative content in Indonesian. The detailed procedures and experimental settings used in this research are described in the following subsections.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

2.1. Dataset Collection

The dataset utilised in this research comprises Indonesian language posts (tweets) gathered from the social media platform X. Because social media allows users to instantly express opinions and reactions to social and political events, it was selected as the main data source. This makes social media a useful tool for analysing public discourse and provocative speech.

Tweets were collected using a crawling method through the X API with the Tweet Harvest library [11], which enables the automatic retrieval of tweets based on predefined search queries. Several keywords related to demonstration events were used to collect relevant tweets discussing the protest activities. The keywords were selected to capture conversations related to demonstrations and potential provocative narratives circulating on social media.

The data were collected within the time range of August 25, 2025, to August 31, 2025, corresponding to the period when the demonstration events occurred. This time frame was chosen to ensure that the collected tweets reflected the context of the ongoing demonstrations and related discussions among social media users.

After the crawling process was completed, a total of approximately 9,500 tweets were obtained. The collected tweets were then filtered to retain only Indonesian-language posts and remove irrelevant content such as advertisements, spam, or duplicated tweets. The resulting dataset was subsequently used for the labeling and preprocessing stages prior to model training.

Table 1 presents the keywords used during the data collection process to retrieve tweets related to the demonstration events.

Table 1 . Keywords Used for Data Collection

No	Keyword	No	Keyword	No	Keyword	No	Keyword
1	demo	11	penghasut	21	aksi	31	lempar
2	gedung dpr	12	kudeta	22	aparat	32	pengkhianat
3	dpr ri	13	rusuh	23	brimob	33	serang
4	aksi demo	14	boikot	24	serbu	34	ganyang
5	turun ke jalan	15	lawan	25	reformasi	35	sebarkan
6	polisi	16	kepung dpr	26	bubarkan	36	boneka
7	tentara	17	datang ke dpr	27	seruan	37	kacung
8	bubarkan dpr	18	bergerak sekarang	28	penjarahan	38	penjilat
9	provokasi	19	bentrok	29	hancurkan	39	menghasut
10	provokator	20	bakar gedung	30	kumpul	40	anarkis

Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1127

Figure 1. WordCloud Provocative Tweets

To provide an overview of the linguistic characteristics of provocative tweets, a word cloud visualization was generated based on tweets labeled as provocative. The visualization highlights frequently occurring words in the dataset, where larger words represent higher frequencies. As shown in Figure 1, several dominant terms related to confrontation, mobilization, and expressions of anger appear frequently. These patterns indicate that provocative tweets often contain emotionally charged language associated with collective actions and public dissatisfaction during the demonstration period.

2.2. Dataset Labelling

After the data collection process, the collected tweets were annotated to determine whether they contained provocative content. The annotation process was carried out by three independent annotators to improve the reliability of the labeling results.

In this research, tweets were categorized into two classes: provocative and non-provocative. The term "provocative" refers to expressions intended to incite or stimulate certain reactions. According to the Indonesian Dictionary (Kamus Besar Bahasa Indonesia), provocative refers to statements or actions intended to provoke or incite particular responses from individuals or groups [3]. Based on this definition, a tweet was labeled as provocative if it contained expressions that encouraged hostility, incited collective action, or provoked public anger toward certain individuals, groups, or institutions. Tweets that only expressed opinions, provided information, or discussed issues without encouraging conflict or mobilization were labeled as non-provocative.

Each annotator labeled the tweets independently. The final label for each tweet was determined using a majority voting scheme, where the label agreed upon by at least two annotators was selected as the final annotation result.

2.3. Text Preprocessing

Text data collected from social media platforms often contain noisy elements such as URLs, mentions, hashtags, emojis, and informal language expressions. Therefore, multiple preprocessing steps were performed to clean and normalize the text before it was used for

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

model training. These preprocessing steps were implemented using Python with regular expression techniques and text processing libraries.

1) Cleaning, the cleaning stage was performed to remove irrelevant elements that do not contribute to the semantic meaning of the text. These elements include URLs (http/https), hashtags (#), mentions (@username), retweet indicators (RT), emojis, and other special characters. In addition, this stage also involves removing numbers, excessive punctuation, and expressions of laughter such as “wkwk”, “haha”, and “hehe”, which are commonly found in social media texts but do not provide meaningful information for classification. Symbol normalization was also applied, for example replacing the symbol “&” with the word “dan” and removing unnecessary symbols such as “>”, “+”, and parentheses. Removing these elements helps reduce noise in the dataset and improves the quality of the input text for the classification model.
2) Case folding, Text normalization includes the application of case folding to convert all text into lowercase, as well as replacing slang words and non-standard spellings with their standard forms using an Indonesian dictionary specifically designed for social media texts (e.g., "gw" is normalized to "saya", and "lu" to "kamu").
3) Stemming, Stemming was performed using the Sastrawi library, which converts affixed words into their base forms to reduce word variations that may confuse the model during training.
4) Tokenization, Tokenization was conducted using the BertTokenizer provided by IndoBERTweet, with a maximum sequence length of 128 tokens. Automatic padding and truncation were applied to ensure that all input sequences have a uniform length.
5) Stopword, Stopword removal was performed to eliminate common function words that carry minimal semantic meaning in the classification process, such as “yang”, “dan”, “atau”, “di”, “ke”, and “dari”. The stopword list used in this research was adapted from the Indonesian stopword list provided by the *stopwords-iso* project [12].

The preprocessing stage was conducted to clean and normalize the collected tweets before further analysis. During this stage, several filtering procedures were applied, including removing duplicate tweets, irrelevant promotional content, and incomplete text entries. As a result, the initial dataset of approximately 9,500 tweets was reduced to 8,899 tweets after the preprocessing process. The cleaned dataset consisted of 5,277 non-provocative tweets and 3,622 provocative tweets, corresponding to approximately 59.3% and 40.7% of the dataset, respectively. The dataset shows a relatively balanced class distribution.

2.4. Model Architecture

This model employs a model architecture based on Bidirectional Encoder Representations from Transformers (BERT), a deep bidirectional language model introduced by Devlin et al. [13]. BERT utilises the transformer encoder architecture to

discern contextual relationships among words by examining text sequences in a bidirectional manner. The BERT-base variant comprises 12 transformer layers, 768 hidden dimensions, and 12 attention heads, yielding approximately 110 million parameters. The model utilises the WordPiece tokeniser for text processing, featuring a vocabulary of approximately 30,000 tokens, and incorporates special tokens like [CLS] and [SEP] for sentence classification and separation. In the pretraining phase, BERT is refined through two primary objectives: Masked Language Modelling (MLM), in which certain tokens are randomly obscured and subsequently predicted, and Next Sentence Prediction (NSP), which assesses whether two sentences are sequentially present in the original corpus. The model was initially trained on extensive English corpora, including BooksCorpus and English Wikipedia, allowing it to acquire comprehensive contextual language representations.

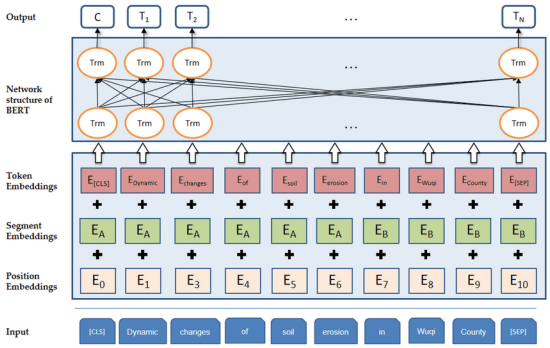

Figure 2. BERT's Structure

To support natural language processing tasks for Indonesian text, Koto et al. [14] introduced IndoBERT, a monolingual adaptation of the BERT architecture specifically trained on Indonesian-language data. IndoBERT maintains the same base architecture as BERT-base, consisting of 12 transformer layers with 768 hidden dimensions and 12 attention heads. The primary difference lies in the training corpus and tokenizer, which are tailored for Indonesian linguistic characteristics. During pretraining, the model was trained on approximately 220 million Indonesian words collected from multiple sources, including Indonesian Wikipedia, online news articles, and web-crawled texts. In addition, IndoBERT uses an Indonesian-specific WordPiece tokenizer with a vocabulary size of 31,923 tokens. By being trained on Indonesian corpora, the model is able to learn semantic and syntactic patterns unique to the language, which helps improve its performance across various downstream natural language processing tasks.

Although IndoBERT performs well on formal Indonesian texts, social media data often contain informal language patterns, slang expressions, abbreviations, and non-standard spellings. To address this challenge, Koto et al. [15] proposed IndoBERTweet, a domain-adapted language model designed specifically for Indonesian Twitter data. IndoBERTweet maintains the same transformer architecture as BERT-base but is pretrained

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

on a large corpus of Indonesian tweets consisting of approximately 26 million tweets with more than 400 million word tokens collected through the Twitter API. The model introduces a new WordPiece vocabulary containing 31,984 tokens, where a significant portion of the vocabulary consists of new tokens specific to social media language. Additional preprocessing techniques are applied to normalize Twitter text, such as replacing user mentions with @USER and URLs with HTTPURL. IndoBERTweet is trained using the Masked Language Modeling objective and adopts a maximum sequence length of 128 tokens to better accommodate the short-text characteristics of tweets. By incorporating domain-specific vocabulary and training data, IndoBERTweet is able to capture the linguistic characteristics of social media discourse more effectively than general Indonesian language models, making it suitable for tasks such as sentiment analysis, hate speech detection, and provocative speech classification on social media platforms.

2.5. Fine Tuning

Fine-tuning is the process of modifying a pre-trained language model for a specific downstream task by adjusting its parameters with labelled data. In transformer-based models like BERT, this procedure is typically executed by incorporating a task-specific classification layer atop the contextual representations produced by the model. The concealed representation of the special token [CLS] is generally employed as the input to the classification layer, subsequently followed by a softmax function to generate class probabilities. This procedure enables the model to leverage linguistic knowledge acquired during the pre-training phase while fine-tuning its parameters for effective performance on a specific task, even with a limited labelled dataset available [13].

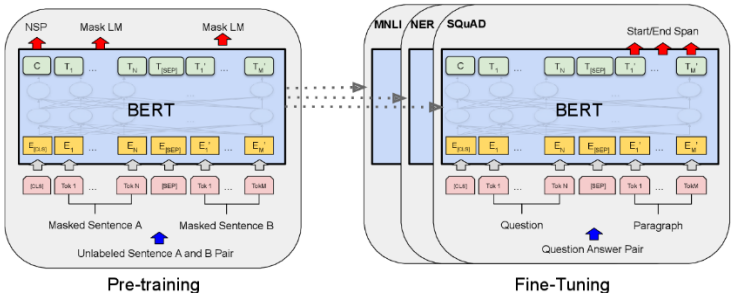

Figure 3. Fine Tuning in BERT

In BERT-based architectures, the fine-tuning process can be conducted either by updating all model parameters or by updating only a subset of them. In the full fine-tuning approach, every parameter in the network is adjusted during the training phase. In contrast, partial fine-tuning freezes several early layers of the model while only the higher layers and the classification head are trained. This strategy helps reduce computational cost and is particularly beneficial when working with small to medium-sized datasets [16].

This research involved fine-tuning the IndoBERTweet model for the detection of provocative speech in Indonesian social media data. IndoBERTweet is a domain-adapted pretrained language model based on IndoBERT, trained on an extensive corpus of Indonesian Twitter data. The model, being specifically trained on social media text, effectively captures informal language patterns, including slang and diverse linguistic variations prevalent in online discussions. IndoBERTweet produces contextual representations tailored for Twitter data analysis through domain-specific pretraining, surpassing general-purpose language models [15].

The fine-tuning procedure in this study was carried out using the Hugging Face Transformers library with PyTorch as the underlying framework. During the training phase, several hyperparameters were determined based on common configurations reported in earlier studies. Previous research on BERT fine-tuning indicates that moderate batch sizes (16–32) and relatively low learning rates contribute to stable training and better model convergence. Additionally, the maximum sequence length is usually limited to reduce computational cost while still preserving sufficient contextual information for the model [17][18].

Furthermore, the learning rate is an important hyperparameter in the fine-tuning process because it controls the magnitude of parameter updates during model training. An inappropriate learning rate may cause unstable training behavior or lead to suboptimal model performance. Previous studies indicate that learning rates within the range of 1×10^{-5} to 5×10^{-5} tend to produce stable results when fine-tuning BERT-based models [13]. Therefore, multiple learning rate settings were evaluated in this study to determine the most effective configuration for provocative speech classification.

Finally, the model was trained using a cross-entropy loss function to distinguish between provocative and non-provocative tweets. Fine-tuning enables the model to transfer knowledge obtained during pre-training on large corpora to the target task, allowing the classifier to achieve strong performance even when the labeled dataset is relatively limited. Previous studies have also shown that fine-tuned BERT-based models generally outperform traditional machine learning approaches such as Naïve Bayes or Support Vector Machine in text classification tasks [19].

2.6. Model Evaluation

The performance of the provocative speech detection model was evaluated using standard classification metrics on the X social media dataset. In this study, the evaluation metrics include accuracy, precision, recall, and F1-score. These metrics are widely applied in text classification research to measure how effectively machine learning and deep learning models perform in predicting the target classes.

Accuracy (the proportion of correctly classified instances among the total number of predictions) (1)

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

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

In this context, TP (True Positive) denotes positive instances accurately classified by the model, TN (True Negative) signifies negative instances correctly predicted, FP (False Positive) represents negative instances erroneously classified as positive, and FN (False Negative) indicates positive instances incorrectly predicted as negative.

Precision (the proportion of true positive predictions among all positive predictions)

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (the proportion of true positive predictions among all actual positive instances)

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-score (the harmonic mean of precision and recall)

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (4)$$

3. RESULTS AND DISCUSSION

This section describes the experimental results and discussion related to the proposed methods for detecting provocative speech on the social media platform X. Several classification approaches were examined, including traditional machine learning techniques, deep learning models, and transformer-based architectures. The performance of each model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

3.1. Model Training Process

This research employed various classification models to detect provocative speech on the social media platform X. The models encompass both traditional machine learning techniques and transformer-based models to facilitate a thorough comparison of their classification efficacy. Conventional machine learning algorithms, including Multinomial Naïve Bayes and Linear Support Vector Machine (SVM), were employed as baseline models owing to their efficacy and prevalent application in text classification tasks. Textual data were transformed into numerical feature representations utilising the Term Frequency–Inverse Document Frequency (TF–IDF) method. Simultaneously, the IndoBERTweet model

underwent fine-tuning utilising the Hugging Face Transformers library, with PyTorch serving as the backend framework. Multiple training hyperparameters were modified to enhance the learning process, encompassing sequence length, batch size, optimiser settings, and learning rate values. The comprehensive hyperparameter configurations employed in this study are delineated in Table 2.

Table 2. Hyperparameter Configuration

Hyperparameter	Value
Model	IndoBERTweet
Maximum sequence length	128
Batch size	16
Optimizer	AdamW
Weight decay	0.01
Learning rate	1e-5, 2e-5, 3e-5, 4e-5, 5e-5
Maximum epoch	2
Early stopping	Patience = 1
Classification threshold	0.45
Validation method	5 – fold cross validation

Various learning rate values were assessed to ascertain the optimal training configuration. An early stopping mechanism was implemented during training to mitigate the risk of overfitting. A classification threshold was employed during the prediction phase to ascertain the final class label.

A 5-fold cross-validation method was utilised to enhance the model's robustness and generalisation capability. This method involved partitioning the dataset into five subsets, with four subsets utilised for model training and the fifth subset designated for validation purposes. The procedure was executed five times, ensuring that each subset served as validation data once. The overall model performance was subsequently determined by averaging the evaluation results derived from all folds.

3.2. Model Evaluation Results

This section presents the evaluation results of the proposed provocative speech detection models. The performance of the models was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

Table 3. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naïve Bayes	83.14%	84.00%	83.00%	83.00%
Linear SVM	89.36%	89.00%	89.00%	89.00%
BiLSTM	88.38%	84.18%	88.25%	86.17%
IndoBERT	92.72%	90.59%	91.66%	91.11%
IndoBERTweet (LR = 5e-5)	92.96%	90.55%	92.35%	91.43%
IndoBERTweet (LR = 4e-5)	93.01%	90.52%	92.55%	91.51%
IndoBERTweet (LR = 3e-5)	**93.07%**	**90.76%**	**92.41%**	**91.56%**
IndoBERTweet (LR = 2e-5)	92.87%	90.56%	92.10%	91.31%
IndoBERTweet (LR = 1e-5)	92.36%	90.01%	91.39%	90.68%

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

The experimental findings indicate that conventional machine learning models typically exhibit inferior performance relative to deep learning and transformer-based approaches, which is consistent with previous studies on Indonesian social media text classification [20]. The Multinomial Naïve Bayes model achieved an accuracy of 83.14% and an F1-score of 83%, whereas the Linear Support Vector Machine (SVM) demonstrated superior performance with an accuracy of 89.36% and an F1-score of 89%. Despite their computational efficiency and widespread application in text classification research, these models exhibit a relatively limited capacity to capture intricate contextual relationships in social media text.

The BiLSTM model, representing a deep learning methodology, exhibited marginally enhanced performance relative to conventional machine learning techniques, attaining an accuracy of 88.38% and an F1-score of 86.17%. Although BiLSTM effectively captures sequential dependencies in textual data, its performance remains inferior to transformer-based models that utilise contextual language representations. This finding is consistent with previous studies showing that deep learning models such as BiLSTM perform effectively in Indonesian social media text classification tasks [21].

Transformer-based models show significantly better performance in detecting provocative speech, which aligns with recent studies and datasets on Indonesian hate speech and social media text classification [22], [23], [24]. The IndoBERT model achieved an accuracy of 92.72% with an F1-score of 91.11%, demonstrating the effectiveness of contextual embeddings for Indonesian text classification. Meanwhile, the IndoBERTweet model achieved the best performance among the evaluated models after hyperparameter tuning. This result indicates that domain-specific language models trained on social media data are more effective in capturing informal linguistic patterns commonly found in online discourse.

This finding is consistent with previous studies that reported the effectiveness of IndoBERTweet for Indonesian Twitter text classification tasks, particularly in hate speech detection, where the model demonstrated strong performance in capturing contextual and informal language patterns [25].

Further experiments were conducted by fine-tuning the IndoBERTweet model using several learning rate configurations to determine the optimal training setting. The results show that the learning rate of 3×10^{-5} produced the best overall performance with an accuracy of 93.07% and an F1-score of 91.56%. Other learning rate values, including 5×10^{-5}, 4×10^{-5}, 2×10^{-5}, and 1×10^{-5}, produced slightly lower results. This finding suggests that selecting an appropriate learning rate plays an important role in the stability of the fine-tuning process for transformer-based models.

Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1127

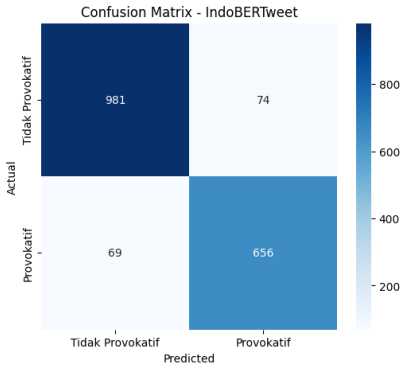

Figure 4. Confusion Matrix of IndoBERTweet

As shown in Figure 4, the model correctly classified 981 non-provocative tweets and 656 provocative tweets, indicating strong capability in distinguishing between the two classes. However, 74 non-provocative tweets were incorrectly classified as provocative, while 69 provocative tweets were misclassified as non-provocative. These misclassifications may occur due to ambiguous language patterns, sarcasm, or contextual nuances commonly found in social media text.

Overall, the results demonstrate that the IndoBERTweet model is able to effectively capture contextual patterns associated with provocative speech in Indonesian social media discourse, while maintaining relatively balanced performance across both classes.

Compared to previous studies that primarily focus on sentiment analysis or hate speech detection in Indonesian social media, this research specifically targets provocative speech within the context of socio-political events. Unlike general-purpose language models, this study leverages IndoBERTweet, which is pretrained on Indonesian Twitter data, allowing it to better capture informal language patterns, slang, and contextual nuances. Therefore, this study provides a more domain-specific and context-aware approach to analyzing provocative discourse, particularly in politically sensitive situations.

4. CONCLUSION

This research examined provocative discourse on the social media platform X during the DPR demonstrations in August 2025 using various text classification approaches. A dataset consisting of 8,899 Indonesian tweets were collected and categorized into two classes: provocative and non-provocative. Several models were evaluated, including traditional machine learning, deep learning, and transformer-based approaches.

Research findings indicate that transformer-based models surpass conventional machine learning techniques in detecting provocative speech in Indonesian social media data. Of the assessed methodologies, IndoBERTweet attained the superior performance, achieving an accuracy of 93.07% and an F1-score of 91.56% with a learning rate of 3 × 10−5.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

The results indicate that domain-specific language models trained on social media corpora are more adept at capturing informal linguistic patterns prevalent in online conversations.

This research contributes both scientifically and practically. From a scientific perspective, it provides a comprehensive comparison of traditional and transformer-based models for detecting provocative speech in Indonesian social media. From a practical perspective, the proposed approach can be applied to monitor and mitigate provocative content on online platforms, particularly during socio-political events. The findings of this study may also support government institutions, digital platform providers, and public communication analysts in identifying potentially provocative discourse and improving content moderation strategies in social media environments.

However, this study has several limitations. The dataset is limited to tweets related to a specific event, which may not fully represent broader public discourse. Additionally, the labeling process relies on manual annotation, which may introduce subjectivity. This study also focuses only on textual data and does not consider multimodal information such as images or videos.

Additional analysis using wordcloud visualization indicates that provocative tweets tend to contain more aggressive and emotionally charged vocabulary. This research contributes to the limited research on provocative speech detection in Indonesian social media by providing a labeled dataset and evaluating multiple classification approaches for identifying provocative discourse during political events. Future research may explore larger datasets, more advanced transformer architectures, and additional contextual features to further improve the detection of provocative speech in Indonesian social media discussions.

REFERENCES

[1] "Viral Demo DPR 25 Agustus 2025, Netizen Ramai Komentar Begini." Accessed: Nov. 17, 2025. [Online]. Available: https://www.cnbcindonesia.com/tech/20250825133814-37-661108/viral-demo-dpr-25-agustus-2025-netizen-ramai-komentar-begini

[2] "Indonesian police clash with students protesting lawmakers' salaries | Protests News | Al Jazeera." Accessed: Apr. 04, 2026. [Online]. Available: https://www.aljazeera.com/news/2025/8/26/indonesian-police-clash-with-students-protesting-lawmakers-salaries

[3] "Arti kata provokatif - Kamus Besar Bahasa Indonesia (KBBI) Online." Accessed: Nov. 18, 2025. [Online]. Available: https://kbbi.web.id/provokatif

[4] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 1175–1181, 2023, doi: 10.14569/IJACSA.2023.01406125.

[5] M. O. Ibrohim 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] H. Rahman, Y. H. Putra, H. Syarif, E. Delenia, Y. Findawati, and D. Purwitasari, "Dangerous Speech in Indonesian Twitter Posts: A Literature Review," *Proceeding - 2024 International Conference on Information Technology Research and Innovation, ICITRI 2024*, pp. 269–274, 2024, doi: 10.1109/ICITRI62858.2024.10698923.

[7] J. Forry Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," *JOIV International Journal on Informatics Visualization*, vol. 7, pp. 773–780, 2023, [Online]. Available: www.joiv.org/index.php/joiv

[8] A. P. J. Dwitama, D. H. Fudholi, and S. Hidayat, "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)," *Jurnal RESTI*, vol. 7, no. 2, pp. 302–309, Apr. 2023, doi: 10.29207/resti.v7i2.4642.

[9] A. R. P. Dewi, S. Riyadi, D. Cahya, N. A. M. Isa, and A. D. Andriyani, "Sentiment Analysis of Pro-Israel Product Boycott Action Using IndoBERT Method on Unbalanced Data," *JUITA: Jurnal Informatika*, vol. 13, pp. 187–197, 2025.

[10] P. Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Journal on Computing*, vol. 9, no. 2, pp. 61–72, 2024, doi: 10.34818/indojc.2024.9.2.934.

[11] "GitHub - helmisatria/tweet-harvest: Scrape tweets from Twitter search results based on keywords and date range using Playwright. Save scraped tweets in a CSV file for easy analysis · GitHub." Accessed: Mar. 09, 2026. [Online]. Available: https://github.com/helmisatria/tweet-harvest

[12] "GitHub - stopwords-iso/stopwords-id: Indonesian stopwords collection · GitHub." Accessed: Mar. 07, 2026. [Online]. Available: https://github.com/stopwords-iso/stopwords-id

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Accessed: Dec. 04, 2025. [Online]. Available: https://www.researchgate.net/publication/328230984_BERT_Pre-training_of_Deep_Bidirectional_Transformers_for_Language_Understanding

[14] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, Nov. 2020, doi: 10.18653/v1/2020.coling-main.66.

[15] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 10660–10668, Sep. 2021, doi: 10.18653/v1/2021.emnlp-main.833.

[16] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 328–339, 2018, doi: 10.18653/V1/P18-1031.

[17] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *Journal of Machine Learning Research*, Oct. 2019, Accessed: Mar. 09, 2026. [Online]. Available: https://arxiv.org/abs/1910.03771v5

[18] S. Eltahier, O. Dawood, and I. Saeed, "BERT Fine-Tuning for Software Requirement Classification: Impact of Model Components and Dataset Size," *Information 2025, Vol. 16, Page 981*, vol. 16, no. 11, p. 981, Nov. 2025, doi: 10.3390/INFO16110981.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

[19] M. Bilal and A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737–2757, Dec. 2023, doi: 10.1007/S10660-022-09560-W.

[20] B. T. Yulianto Darmawan, B. R. Irnawan, and Y. Suzuki, "Indonesian Hate Speech and Abusive Tweets Classification with Deep Learning Pre-trained Language Models," *Proceedings - 2023 6th International Conference on Computer and Informatics Engineering: AI Trust, Risk and Security Management (AI Trism), IC2IE 2023*, pp. 30–35, 2023, doi: 10.1109/IC2IE60547.2023.10331354.

[21] R. A. Saputra and Y. Sibaroni, "Multilabel Hate Speech Classification in Indonesian Political Discourse on X using Combined Deep Learning Models with Considering Sentence Length," *Jurnal Ilmu Komputer dan Informasi*, vol. 18, no. 1, pp. 113–125, Feb. 2025, doi: 10.21609/JIKI.V18I1.1440.

[22] M. I. Wijanarko, L. Susanto, P. A. Pratama, I. Idris, T. Hong, and D. Wijaya, "Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts," *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of System Demonstrations*, pp. 142–152, 2024, doi: 10.18653/V1/2024.EMNLP-DEMO.15.

[23] L. Susanto *et al.*, "IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language," Jun. 2024, Accessed: Apr. 04, 2026. [Online]. Available: https://arxiv.org/pdf/2406.19349

[24] E. F. Cahyani, A. Nur Ikhsan, D. N. Astrida, and A. N. Ikhsan, "Event-Based Detection of Provocative Political Discourse on Indonesian Twitter: A Comparative Study of SVM and IndoBERT," *Journal of Information Systems and Informatics*, vol. 8, no. 1, pp. 530–548, Feb. 2026, doi: 10.63158/JOURNALISI.V8I1.1409.

[25] P. A. Mufva, K. H. Chandra, K. F. Aji, I. A. Iswanto, and S. Joddy, "Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding," *Procedia Comput. Sci.*, vol. 269, pp. 1663–1671, 2025, doi: 10.1016/J.PROCS.2025.09.109.