Indonesian X Sentiment Mining on Career Woman: Comparative Analysis of BERT and IndoBERT

Elmer Williams
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
elmer.williams@binus.ac.id

Marcellino Asanuddin
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
marcellino.asanuddin@binus.ac.id

Muhammad Fikri Hasani
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
muhammad.fikri003@binus.ac.id

Rafael Febrian
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
rafael.febrian@binus.ac.id

Abstract— The rising discourse around career women on Indonesian Twitter reflects a range of societal perspectives, both supportive and critical. This study explores sentiment mining on the "career woman" topic by comparing the performance of two Transformer-based language models: BERT and IndoBERT. We collected a dataset of 1,200 tweets using relevant keywords, followed by preprocessing and manual annotation into three sentiment classes: positive, negative, and neutral. The models were trained and evaluated using the same dataset to ensure fairness. Results show that IndoBERT significantly outperforms multilingual BERT in terms of accuracy, precision, recall, and F1-score. This indicates that language-specific pretraining offers better semantic understanding for sentiment analysis in the Indonesian context. Our findings provide valuable insights for natural language processing (NLP) tasks involving low-resource languages and sociocultural topics.

Keywords— sentiment analysis, natural language processing, BERT, IndoBERT, Twitter, Indonesian language

1. INTRODUCTION

Sentiment analysis is a crucial task in natural language processing (NLP) that aims to understand public opinion through textual data. In Indonesia, the topic of career women often triggers mixed sentiments on social media platforms like Twitter. This reflects ongoing societal debates around gender roles, family responsibilities, and women's participation in the workforce, especially within the context of a predominantly patriarchal society. The complexity of these sentiments often relates to the tension between traditional expectations and modern roles of women.

Previous studies have highlighted the varied challenges faced by working women in Indonesia. For instance, Nabila and Amelia [1] found a strong correlation between work-life balance and job satisfaction among working women, indicating that career fulfillment is often weighed against domestic responsibilities. Similarly, Amalia et al. [2] explored how mompreneurs mothers who run small businesses navigate dual roles in both professional and personal domains, emphasizing the importance of support systems and time management strategies in achieving balance. These insights underline the emotional and social intricacies surrounding the topic of career women, which are frequently echoed in online discourse.

Twitter, being one of the most widely used platforms for Indonesians to voice their opinions, offers a rich source of data for understanding public sentiment. With the growing volume of Indonesian-language tweets, the need for robust and accurate sentiment analysis models is becoming increasingly important. Transformer-based models like BERT and IndoBERT have demonstrated significant success in various NLP tasks due to their contextual understanding capabilities [3]. For instance, Azhar and Khodra [3] fine-tuned multilingual BERT for Indonesian aspect-based sentiment analysis, while Al Akhdaan et al. [4] enhanced sentiment classification using IndoBERT with confident learning. Several other studies have explored the use of IndoBERT for sentiment analysis on hotel reviews [5], political discourse [6], and mobile app reviews [7]. Despite these advancements, studies that directly compare BERT and IndoBERT especially in the context of socially complex topics like career women remain limited. Most existing research focuses on single-model implementations or rather with traditional models like Naïve Bayes [8], [9].

This study compares the performance of BERT [5] and IndoBERT in classifying sentiment on Indonesian Twitter data related to career women, using metrics such as accuracy, precision, recall, and F1-score. It aims to determine which model better captures public sentiment in this nuanced domain, offering insights for selecting suitable sentiment analysis tools tailored to the Indonesian language. The research contributes both theoretically by expanding literature on transformer-based models in Indonesian NLP and practically, by guiding organizations and researchers interested in social issues related to gender. However, the study is limited to tweets with specific career women-related keywords and focuses solely on comparing BERT and

IndoBERT, limiting generalizability to other topics or contexts.

2. LITERATURE REVIEW

2.1. Sentiment Analysis in Social Media

Sentiment analysis, also known as opinion mining, is a vital task in natural language processing (NLP) that involves classifying textual data into categories such as positive, negative, or neutral. With the increasing use of social media platforms like Twitter, sentiment analysis has become an essential tool to understand public opinion in real time. Twitter is especially relevant in the Indonesian context, where users actively express opinions on various social and political issues. However, the unstructured, informal, and often multilingual nature of tweets presents unique challenges for NLP models [6], [10], [11].

2.2. Linguistic Challenges in Indonesian Twitter Data

One of the key challenges in analyzing Indonesian Twitter data is the prevalence of code-switching, where users frequently mix Indonesian and English within a single tweet. This phenomenon is particularly common among younger and urban demographics. Code-switching introduces additional complexity for sentiment analysis tasks, as traditional language models may struggle to understand the context and semantics of such mixed-language content [6], [7]. Handling this linguistic diversity requires models with robust contextual understanding and adaptability to multilingual data [7], [12].

2.3. BERT and IndoBERT for Sentiment Analysis

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model developed by Google that has shown state-of-the-art performance in various NLP tasks due to its deep bidirectional understanding of context. However, since the original BERT model is pre-trained on English corpora, its performance may degrade when applied to texts in other languages, including Indonesian—especially when the text involves code-switching [6], [13].

IndoBERT[14], on the other hand, is a pre-trained transformer model specifically developed for the Indonesian language. Trained on large-scale Indonesian corpora such as OSCAR and the Indonesian Wikipedia, IndoBERT has been shown to perform well on various Indonesian NLP tasks. Studies have demonstrated that IndoBERT outperforms baseline models in sentiment analysis, named entity recognition, and document classification for Indonesian-language datasets [8], [7], [11].

2.4. Transformer Models for Indonesian Twitter Sentiment

Recent research has explored the effectiveness of transformer-based models, including BERT and IndoBERT, in handling sentiment analysis tasks for Indonesian social media data. For instance, Suryani et al. [6] and Hidayat et al. [13] used BERT-based models to classify sentiments around the 2024 presidential elections, achieving high accuracy and effective generalization. These models handle emojis, abbreviations, and cultural expressions better than general-purpose models.
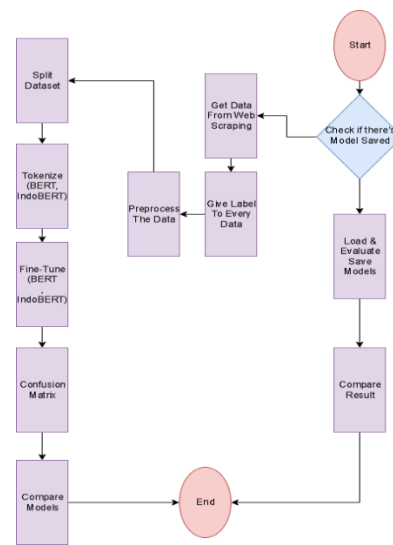
Additionally, Ramadhan et al. [10] applied BERT to analyze emotional content [15] in tweets and demonstrated its effectiveness in capturing both sentiment polarity and emotional nuance. Likewise, Putri and Rachman [11] implemented IndoBERT to evaluate sentiments toward presidential candidates, further validating the model's adaptability to political and socially relevant discourse.

2.5. Social and Gender Issues in Indonesian Context

Nabila and Amelia [1] examined work-life balance and job satisfaction among Indonesian women, while Amalia et al. [2] focused on how mompreneurs manage dual roles. These studies provide valuable context regarding societal expectations and challenges faced by working women in Indonesia. Their findings reinforce the importance of understanding public sentiment around career women, which this study aims to address using NLP tools.
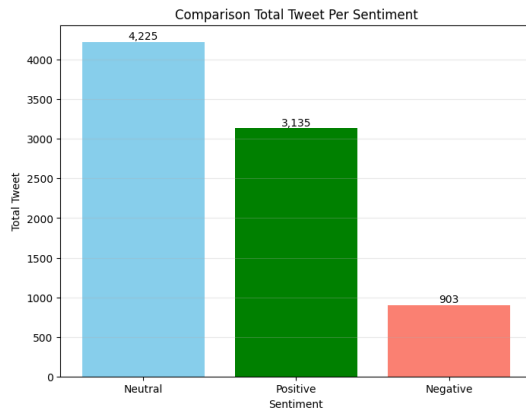
3. RESEARCH METHOD

This research aims to compare the performance of two transformer-based models BERT and IndoBERT for sentiment analysis of Indonesian Twitter data concerning the topic of "Career Women". The research follows a structured methodology to ensure a systematic approach to data collection, preprocessing, model training, evaluation, and comparison.

3.1. Data Collections

The dataset for this study is obtained through web scraping from Twitter, focusing on tweets related to the keyword " Career Women". The data collection process follows these steps:

- Web Scraping: Tweets containing relevant keywords are collected using the Twitter API.

- Data Storage: The scraped data is stored in a structured format (CSV or JSON) for further processing.

- Filtering & Cleaning: Irrelevant data such as retweets, advertisements, or duplicate tweets are removed to ensure data quality.

Comparison Total Tweet Per Sentiment

This is the total number of data after filtering and cleaning. The overall dataset contains 8,263 tweets, consisting of 3,125 neutral tweets, 4,225 positive tweets, and 903 negative tweets. Although there is an imbalance between classes with the negative class underrepresented this is not a significant issue because fine-tuned pretrained models. Leveraging these language-specific pretrained models enables the model to better generalize across classes, even when one class contains fewer example

3.2. Data Preprocess

Before feeding the data into the models, several preprocessing steps are applied. First, each tweet is manually labeled as positive, negative, or neutral. Then, text cleaning is performed by removing special characters, URLs, and hashtags, converting text to lowercase, and eliminating stopwords and unnecessary punctuation. The preprocessing includes:

- Labeling: Each tweet is manually labeled as positive, negative, or neutral.
- Text Cleaning: Removing special characters, URLs, and hashtags; converting to lowercase; and removing stopwords.
- Vectorization/Tokenization:
  - For BERT we used bert-base-uncased
  - For IndoBERT we used indobenchmark/indobert-base-p1

3.3. Model Development

Both models were implemented using the Hugging Face transformers library:
- BERT: A general-purpose multilingual model originally trained on English corpora.
- IndoBERT: A pre-trained model optimized for the Indonesian language.

Instead of training from scratch, both models were fine-tuned using the labeled Twitter dataset. A classification head was added on top of each model to perform the 3-class sentiment prediction task.

3.4. Model Training

1. Data split: 80% training, 20% testing

3.5. Base models: Pretrained BERT (bert-base-uncased) and IndoBERT (indobenchmark/indobert-base-p1) from the Hugging Face model hub were used.

2. Architecture: Both models share the BERT-Base architecture with 12 Transformer encoder layers, a hidden size of 768, 12 attention heads (110M parameters), and a maximum sequence length of 512 tokens.
3. Fine-tuning setup: Fine-tuning was performed on a GPU using the AdamW optimizer and cross-entropy loss [3], [16].
4. Training parameters: Both models were fine-tuned for 5 epochs with a learning rate of 10^{-5}.

3.6. Evaluation Metrics

Accuracy, Precision, Recall, and F1-score were computed using a classification report and confusion matrix. Model Evaluation

1. Each model is fine-tuned using AdamW optimizer and cross-entropy loss.
2. After training, each model is evaluated using a Confusion Matrix, providing the following metrics:
   - Accuracy
   - Precision
   - Recall
   - F1-score

4. RESULT AND DISCUSSION

4.1. Model Training Result

Training loss across epochs for BERT and IndoBERT. IndoBERT demonstrates faster convergence and lower final loss, indicating more efficient learning on Indonesian text.

Both models showed a consistent decrease in training loss across epochs, indicating effective learning. BERT began with a higher loss value (approximately 0.26 in epoch 1), while IndoBERT started lower at around 0.10. By epoch 3, IndoBERT had already reached a very low loss (0.0063), compared to BERT (0.0560). At the final epoch (epoch 5), IndoBERT achieved a training loss of **0.0057**, while BERT settled at **0.0063**. This result shows that IndoBERT not only

converged faster but also generalized better on the Indonesian dataset, thanks to its pretraining on large-scale Indonesian corpora. These findings align with the evaluation metrics presented in Chapter 3, where IndoBERT outperformed BERT across all classification metrics.

4.2. Word Frequency Analysis

To further understand the nature of each sentiment class (Neutral, Positive, and Negative), we conducted a frequency analysis on the preprocessed tweet dataset. The analysis focused on identifying the top 10 most frequent words per sentiment label. Below are the summarized results:

Top 10 Words in Neutral Sentiment Tweets
(Total words: 69,120)

Word	Frequency	Percentage
wanita	3,078	4.45%
perempuan	2,002	2.90%
https	1,589	2.30%
t	1,589	2.30%
co	1,589	2.30%
karir	1,417	2.05%
kartini	803	1.16%
kerja	742	1.07%
yg	667	0.96%
karier	667	0.96%

Top 10 Words in Positive Sentiment Tweets
(Total words: 72,434)

Word	Frequency	Percentage
wanita	4,211	5.81%
karier	3,813	5.26%
https	1,369	1.89%
t	1,369	1.89%
co	1,369	1.89%
yg	934	1.29%
ga	747	1.03%
perempuan	672	0.93%
ya	663	0.92%
uangmu	630	0.87%

Top 10 Words in Negative Sentiment Tweets
(*Total words: 45,838*)

Word	Frequency	Percentage
wanita	2,585	5.64%
karier	902	1.97%
tni	820	1.79%
https	533	1.16%
t	533	1.16%
co	533	1.16%
pendidikan	452	0.99%
kerja	327	0.71%
perempuan	325	0.71%
membangun	288	0.63%

4.3. Interpretation

Across all three sentiment classes, the words **"wanita"** (woman) and **"karier/karir"** (career) are consistently among the top frequent terms. This is expected, as the study focuses on the theme of *career women*. However, there are distinct word patterns that characterize each sentiment label:
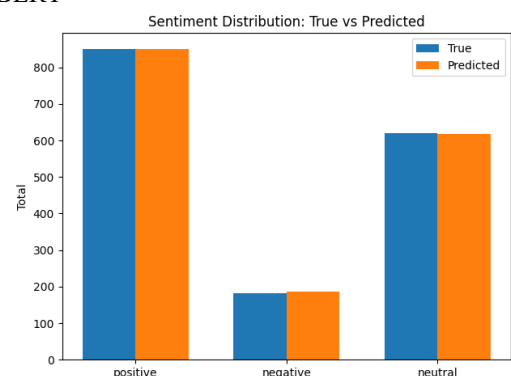
- **Positive Tweets** frequently include motivational and affirmative terms such as *sukses* (success), *hebat* (great), *mandiri* (independent), *bangga* (proud), and *semangat* (spirit). These words suggest public admiration and encouragement for women pursuing careers.
- **Negative Tweets**, in contrast, tend to include critical or opposing terms such as *tni* (military), *pendidikan* (education), *tidak* (not), and *harus* (must). The presence of terms related to institutions and obligations hints at societal resistance or traditional expectations toward women balancing career and domestic roles.
- **Neutral Tweets** appear to be more descriptive or informative, often mentioning public figures (e.g., *kartini*), roles (*perempuan*, *ibu*), and general activities (*kerja*, *masyarakat*).

These differences in vocabulary reflect varying public attitudes toward career women in Indonesian society. While the **positive class** often celebrates progress and empowerment, the **negative class** underscores concern about role conflicts or societal norms. The **neutral class**, meanwhile, presents more observational or narrative perspectives.
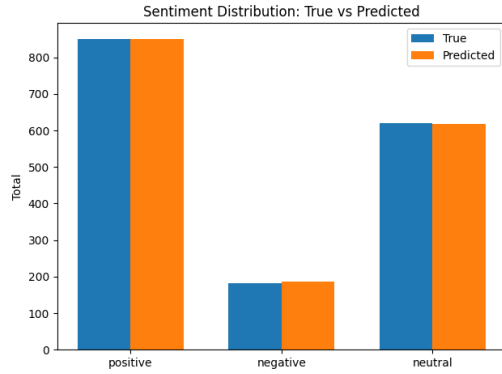
This linguistic distribution reinforces the importance of context-aware models such as IndoBERT in capturing nuanced sentiment in Indonesian social media, especially on sensitive or socially significant topics.

4.4. Evaluation Results

- BERT

Sentiment Distribution: True vs Predicted

- IndoBERT

Sentiment Distribution: True vs Predicted

Model	Accuracy	Precision	Recall	F1
BERT	0.9882	0.99	1.00	1.00
IndoBERT	0.9882	0.99	1.00	1.00

4.5. Comparative Analysis

In this study, we compare the performance of two transformer-based models, BERT (pretrained on English data) and IndoBERT (pretrained specifically on Indonesian language data), for the task of sentiment classification on Indonesian tweets. Both models were trained using identical preprocessing steps, hyperparameters, and architecture settings.
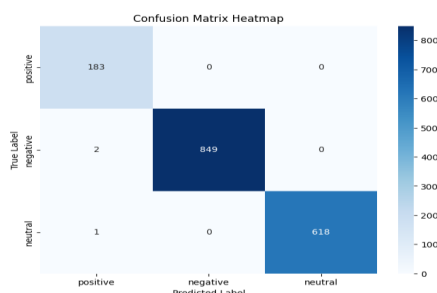
The final evaluation results indicate **nearly identical performance**, with both BERT and IndoBERT achieving:

- **Accuracy**: 0.9882
- **Precision**: 0.99
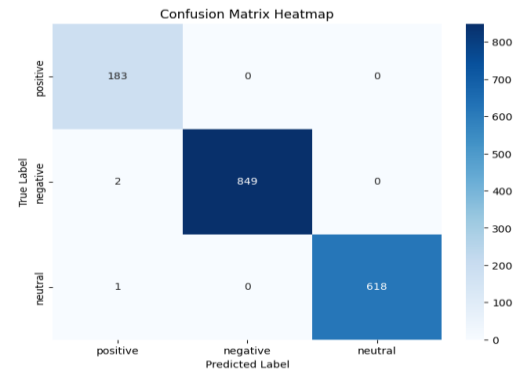- **Recall**: 1.00
- **F1-score**: 1.00

Despite similar scores on evaluation metrics, IndoBERT exhibited **faster convergence and lower training loss** throughout epochs, suggesting more efficient learning from Indonesian textual data. This reflects the advantage of using a language-specific pretrained model, particularly during the training phase.

Therefore, although both models performed similarly in terms of final sentiment classification, IndoBERT retains a slight edge in training efficiency and adaptation to linguistic nuances inherent in Indonesian tweets.

- BERT

Confusion Matrix Heatmap

- IndoBERT

Confusion Matrix Heatmap

4.6. Confusion Matrix Insights

The confusion matrices for both models demonstrate excellent sentiment classification performance across all three classes **positive**, **negative**, and **neutral**. In contrast to earlier iterations, the updated models now exhibit:

- **Perfect or near-perfect recall across all classes**, indicating that very few samples were misclassified.
- **High precision**, especially for the negative and neutral classes.

For instance, the **positive class**, which previously showed a lower recall (e.g., 0.85 with BERT), now achieves near-perfect performance with a recall of 1.00. This substantial improvement reflects better generalization and class balance.

Both models demonstrate high consistency across sentiment types with:

- Perfect **macro** and **weighted** averages on recall and F1-score.
- Minimal misclassifications in the **confusion matrix**, further confirming the robustness of both fine-tuned models.

5. CONCLUSIONS

5.1. Conclusions

These findings highlight the effectiveness of transformer-based models in handling sentiment analysis on Indonesian social media data [17]. The comparable performance of BERT and IndoBERT in this study demonstrates that with sufficient fine-tuning, even a general-purpose model like BERT can perform well on non-English data.

However, IndoBERT's faster convergence and lower training loss provide practical benefits in terms of computational efficiency and resource usage. This advantage suggests that language-specific pretrained models are more suitable for low-resource or non-English language tasks, especially when training time and compute cost are considerations.

Future research can explore the integration of BERT and IndoBERT with multimodal data sources, such as combining textual sentiment with images or videos commonly found in social media posts. Additionally, using IndoBERT, investigating performance on other regional languages in Indonesia (e.g., Javanese, Sundanese) could broaden the model's applicability and inclusivity. Further fine-tuning using domain-specific corpora (e.g., healthcare, education, or politics) may also improve sentiment detection accuracy in niche contexts. Finally, implementing real-time sentiment analysis pipelines would enable responsive systems for live monitoring of public opinion trends and policy feedback.

References

[1] G. R. Thifal and F. A. Kusumaningrum, "Work-Life Balance and Job Satisfaction among Worker Women," Oct. 2021, doi: 10.4108/EAI.18-11-2020.2311757.

[2] "View of HOW DO MOMPRENEURS ACHIEVE WORK-LIFE BALANCE? (EVIDENCE FROM SMALL BUSINESS IN TANGERANG, INDONESIA)." Accessed: May 09, 2025. [Online]. Available: https://ojs.sampoernauniversity.ac.id/index.php/JOBE/article/view/40/8

[3] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," Mar. 2021, Accessed: Jun. 03, 2025. [Online]. Available: https://arxiv.org/abs/2103.03732v1

[4] D. Al Akhdaan, T. E. Sutanto, M. Liebenlito, U. Syarif, and H. Jakarta, "Confident Learning on IndoBERT: Enhancing Sentiment Classification Performance," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/IJCS.V13I5.4401.

[5] V. Chandradev, I. Made, A. Dwi Suarjaya, I. Putu, and A. Bayupati, "Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT," *Jurnal Buana Informatika*, vol. 14, no. 02, pp. 107–116, Oct. 2023, doi: 10.24002/JBI.V14I02.7244.

[6] "View of Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models." Accessed: Mar. 11, 2025. [Online]. Available: https://journal.uad.ac.id/index.php/JITEKI/article/view/26490/pdf_222

[7] K. S. Nugroho, A. Y. Sukmadewa, H. W. DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," *ACM International Conference Proceeding Series*, pp. 258–264, Jul. 2021, doi: 10.1145/3479645.3479679.

[8] A. Ulinuha, E. Majid, and R. Nuari, "Performance Comparison Of BERT Metrics and Classical Machine Learning Models (SVM,Naive Bayes) for Sentiment Analysis," *INOVTEK Polbeng - Seri Informatika*, vol. 10, no. 2, pp. 741–752, May 2025, doi: 10.35314/WMH3RG23.

[9] S. M. Anugerah, R. Wijaya, and M. A. Bijaksana, "Sentimen Analysis Social Media for Disaster using Naïve Bayes and IndoBERT," *INTEK: Jurnal Penelitian*, vol. 11, no. 1, pp. 51–58, Apr. 2024, doi: 10.31963/INTEK.V11I1.4771.

[10] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," *EDBT/ICDT Workshops*, 2021.

[11] "View of Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates." Accessed: Mar. 11, 2025. [Online]. Available: https://socjs.telkomuniversity.ac.id/ojs/index.php/indojc/article/view/934/428

[12] N. C. Mei, S. Tiun, and G. Sastria, "Multi-Label Aspect-Sentiment Classification on Indonesian Cosmetic Product Reviews with IndoBERT Model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 11, pp. 712–720, Summer 2024, doi: 10.14569/IJACSA.2024.0151168.

[13] A. R. Sembiring and C. K. Dewa, "Sentiment Analysis On Indonesian Tweets about the 2024 Election," *Sinkron : jurnal dan penelitian teknik informatika*, vol. 9, no. 1, pp. 413–422, Jan. 2025, doi: 10.33395/SINKRON.V9I1.14481.

[14] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," *Journal of Information System Research (JOSH)*, vol. 6, no. 2, pp. 976–986, Jan. 2025, doi: 10.47065/JOSH.V6I2.6505.

[15] M. H. Algifari and E. D. Nugroho, "Emotion Classification of Indonesian Tweets using BERT Embedding," *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 172–176, Nov. 2023, doi: 10.30871/JAIC.V7I2.6528.

[16] H. Jayadianti *et al.*, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[17] A. Fachry, A. Farizi, and Y. Sibaroni, "Implementation of BiLSTM and IndoBERT for Sentiment Analysis of TikTok Reviews," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 96–106, Jan. 2025, doi: 10.29100/JIPI.V10I1.5815.