IJSA (2024), Vol. 8, No. 2, 83-94
DOI: https://doi.org/10.29244/ijsa.v8i2p83-94

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Sentiment Classification on the 2024 Indonesian Presidential Candidate Dataset Using Deep Learning Approaches*

Cici Suhaeni[1‡], Hari Wijayanto[2], Anang Kurnia[3]

[1,2,3]The Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Indonesia
‡corresponding author: cici_suhaeni@apps.ipb.ac.id

Abstract

This study aims to compare the performance of three deep learning models (LSTM, BiLSTM, and GRU) in the task of sentiment classification for the 2024 Indonesian Presidential Candidate dataset, focusing specifically on the case of Prabowo Subianto. The dataset comprises social media X posts sourced from kaggle, and the analysis investigates the effectiveness of different variants of recurrent neural network architectures in identifying public sentiment. The models were evaluated on accuracy and F1 score. The results demonstrate that BiLSTM outperformed both LSTM and GRU models in all metrics, achieving a testing accuracy of 80.70% and an F1 score of 86.86%, compared to LSTM and GRU which both achieved a testing accuracy of 72.56% and an F1 score of approximately 84%. The higher performance of BiLSTM is attributed to its ability to capture bidirectional context within the text, thereby understanding complex sentiment patterns more effectively. LSTM and GRU models displayed similar performance, therefore BiLSTM is the best model for this dataset. These results indicate that BiLSTM is especially well-suited for analyzing public sentiment towards political figures like Prabowo Subianto, offering significant insights into public discussions surrounding the 2024 Indonesian Presidential Election. This study recommends exploring transformer-based models like BERT or GPT variants to enhance sentiment classification accuracy in this domain.

Keywords: Deep Learning, GRU, Indonesian Presidential Election, LSTM, Sentiment Analysis.

* Received: Dec 2024; Reviewed: Dec 2024; Published: Dec 2024

1. Introduction

The widespread use of social media has transformed it into a popular platform for discussing political issues. Platforms like Twitter (now rebranded as X) are commonly used by the public to express their views and opinions about political events and candidates. The growing influence of these platforms, especially during electoral periods, has made them valuable sources of data for understanding public sentiment towards political figures (Tumasjan et al., 2010).

Social media plays a crucial role in generating political awareness and fostering public discussion, enabling individuals to express their opinions and actively participate in political processes. It serves as an effective communication platform, allowing stakeholders to better understand and respond to public engagement during elections (Adams et al., 2024) . By leveraging this platform, sentiment analysis can be applied to social media data to quantitatively assess public sentiment, providing deeper insights into the reception of political messages and the evolving dynamics of voter sentiment over time. In the context of the 2024 Indonesian Presidential Election, social media has emerged as a critical space for public discourse, making it essential to analyze the sentiment expressed about key candidates like Prabowo Subianto to better understand public opinion trends.

Sentiment analysis has become an essential instrument across various domains, particularly for political analysis, forecasting events, and policy-making (Islam et al., 2024). Its applications extend to fields such as e-commerce, healthcare, education, and social media, with deep learning approaches proving more effective than traditional machine learning and lexicon-based methods (Firdaus et al., 2024). Additionally, sentiment analysis has made significant contributions to understanding global events, such as the COVID-19 pandemic (Abiola et al., 2023), and identifying offensive content on social platforms (Bonetti et al., 2023), as well as analyzing reviews involving imbalanced datasets (Suhaeni & Yong, 2023, 2024).

The rapid advancement of sentiment analysis is reflected in a substantial volume of literature examining its development and applications (Hartmann et al., 2023). Reviews have provided comprehensive insights into sentiment analysis trends (Bordoloi & Biswas, 2023; Sahoo et al., 2023; Tan et al., 2023), and some studies specifically focus on sentiment analysis using Twitter data (Singh & Kumar, 2023). In the political domain, (Ansari et al., 2020) utilized Long Short-Term Memory (LSTM) and Random Forest models to investigate political sentiments on Twitter, showing favorable outcomes for both. Hananto et al. (2023) analyzed trends in presidential candidates using sentiment analysis by comparing three traditional machine learning algorithms, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Naïve Bayes (NB), highlighting their comparative effectiveness for the Indonesian presidential elections. More recently, Ma'aly et al., 2024 analyzed comments on YouTube videos related to the 2024 Indonesian presidential debates by employing Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and a hybrid CNN-BiLSTM architecture, concluding that the BiLSTM model yielded the highest accuracy.

LSTM is a variant of Recurrent Neural Networks (RNN) designed to store

information over extended periods, thereby enhancing classification performance by capturing long-range dependencies in the data. LSTMs employ advanced memory mechanisms that can read, write, and forget information, effectively addressing the vanishing gradient problem often found in traditional RNNs (Ragheb et al., 2019) . BiLSTM (Bidirectional LSTM) extends LSTM capabilities by processing input data in both forward and backward directions, allowing for a more comprehensive understanding of context, which makes it especially effective in sentiment analysis tasks where context plays a critical role (Ragheb et al., 2019; Schuster & Paliwal, 1997). On the other hand, Gated Recurrent Unit (GRU) is a simplified version of LSTM, with fewer parameters, resulting in faster training times while maintaining the ability to learn long-term dependencies (Cho et al., 2014; Han et al., 2021). GRUs employ two main gating mechanisms (reset and update gates) to determine how much of the past information should be retained or discarded, helping maintain efficient training (Han et al., 2021). These characteristics make LSTM, BiLSTM, and GRU well-suited for handling sequence data, although each model offers distinct advantages depending on the specific requirements of the analysis.

Given the strengths and unique features of these deep learning models, selecting the most effective approach for a specific social media dataset becomes crucial. Therefore, the primary objective of this study is to compare the performance of three deep learning models (LSTM, BiLSTM, and GRU) in the sentiment classification task for the 2024 Indonesian Presidential Candidate dataset. Specifically, the study aims to evaluate how well each model performs in identifying public sentiment regarding Prabowo Subianto by focusing on key performance metrics such as accuracy and F1 score. This research intends to provide insights into the most suitable model for sentiment analysis in the political domain, particularly for understanding public discourse surrounding electoral candidates based on this specific dataset.

2. Metodology

2.1 Data

The dataset used in this study was sourced from Kaggle, titled "Indonesia Presidential Candidate's Dataset, 2024" (Dumlao, 2024). This analysis specifically focuses on tweets related to Prabowo Subianto, collected between December 2022 and April 2023, before the 2024 Indonesian presidential election. The data, which originated from Twitter (now rebranded as X), was obtained using Python and the Twitter API.

The original dataset consisted of 9,913 tweets discussing issues related to Prabowo Subianto. After first step of preprocessing, which included removing missing values and duplicate entries, the dataset was refined to 6,757 tweets. This cleaned dataset was used for the sentiment analysis in this study.

Initially, the tweets were in Indonesian. However, the dataset owner translated them into English and labeled them with sentiment categories: Positive or Negative. Thus, the dataset used in this study is the English-translated version, complete with sentiment labels, and ready for further analysis.

2.2 The Data Analysis

The data analysis process involved several stages to prepare, train, and evaluate the models used for sentiment classification. Below, each step is described in detail:

1) Data Preprocessing

The first step in data analysis was preprocessing to ensure that the dataset was clean and consistent for modeling. The preprocessing involved:

- Removing Missing Values and Duplicate Data: Initially, the dataset was cleaned by removing missing values and duplicate entries, reducing the dataset to 6,757 tweets.
- Lowercasing: All text was converted to lowercase to maintain case consistency, ensuring that words like "Indonesia" and "indonesia" were treated as the same token.
- Removing Punctuation and Non-Alphabetic Characters: Punctuation marks and non-alphabetic characters were removed, leaving only alphabetic characters. This helped simplify the text for analysis.
- Removing Extra Spaces: Extra spaces were removed to ensure that the text was formatted cleanly, preventing issues during tokenization.

2) Label Encoding

Sentiment labels in the dataset (Positive and Negative) were encoded into numerical values to be used by the machine learning models. The label encoder transformed the labels into binary values: 0 for Negative and 1 for Positive.

3) Splitting the Data

The dataset was divided into three sets: training, validation, and testing, to evaluate model performance effectively:

- Training Set: 60% of the data was used for training the model.
- Validation Set: 20% of the data was used for validation during training to tune hyperparameters and prevent overfitting.
- Testing Set: 20% of the data was reserved for final testing to evaluate the model's performance on unseen data.

The data was split in two stages: initially, 60% was allocated for training, and the remaining 40% was split evenly between validation and testing sets.

4) Pre-Sentiment Classification

The preparation phase for sentiment classification involved two key steps to convert the textual data into a format suitable for deep learning models:

- Tokenization: The text data was tokenized using the Keras Tokenizer. This involved converting words into sequences of integers, with a vocabulary size of 5,000 and an out-of-vocabulary token (<OOV>) for unknown words. Tokenization transformed the text into a numerical representation suitable for model input.
- Sequence Padding: To ensure uniform input sizes, all sequences were padded or truncated to a maximum length of 100 tokens. This was necessary for efficient processing by the deep learning models, which require inputs of consistent length. To ensure uniform input sizes, all sequences were padded

or truncated to a maximum length of 100 tokens. This was necessary for efficient processing by the deep learning models, which require inputs of consistent length.

5) Sentiment Classification

For the sentiment classification task, three different deep learning models were implemented and evaluated: LSTM, BiLSTM, and GRU. Each model followed a similar architecture, allowing for a fair comparison of their capabilities in sentiment classification.

Key differences in how the models work are explained below:

- LSTM: Utilizes memory cells with input, output, and forget gates to capture long-term dependencies in sequential data. It processes the data in a single direction (forward).
- BiLSTM: An extension of LSTM that processes the input sequence in both forward and backward directions. This allows the model to capture context from both past and future sequences, improving performance in tasks where bidirectional context is important.
- GRU: A simpler variant of LSTM that combines the functionality of input and forget gates into a single update gate, and lacks a separate memory cell. This makes GRU computationally more efficient, while still capturing dependencies in sequential data effectively.

The architecture included the following components:

- Embedding Layer: Input dimension of 5,000, output dimension of 64, input length of 100.
- Recurrent Layers: Depending on the model, these were LSTM, BiLSTM, or GRU layers with configurations as follows:
  - LSTM Model: Two LSTM layers, with the first containing 64 units and returning sequences, followed by a second LSTM layer with 32 units.
  - BiLSTM Model: Two Bidirectional LSTM layers, with the first containing 64 units and returning sequences, followed by a second layer with 32 units.
  - GRU Model: Two GRU layers, with the first containing 64 units and returning sequences, followed by a second GRU layer with 32 units.
- Dropout Layer: A dropout rate of 0.5 was applied after the first recurrent layer to prevent overfitting.
- Dense Layer: The final dense layer used a sigmoid activation function for binary sentiment classification.

All models were compiled using the Adam optimizer, with binary cross-entropy as the loss function, and accuracy as the evaluation metric. Training was conducted over 5 epochs with a batch size of 64, utilizing the training set, and validated on the validation set. The choice of 5 epochs is not an absolute rule; rather, it is based on common practices and supported by preliminary experiments conducted in this study, which indicated that 5 epochs provided the most optimal performance.

6) Evaluation

The models were evaluated on the test set to determine their effectiveness in sentiment classification. We used accuracy and F1-score as defined by (Terven et al., 2024) to evaluate the classification performance. Accuracy measures the proportion of samples correctly classified compared to the total number of samples, with the formula:

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

With TP= true positive,TN = true negative, FP=false positive, and FN=false negative. While a higher accuracy indicates better performance, it may not always represent the best model, especially in imbalanced datasets. On the other hand, F1-score combines precision and recall into a single metric to evaluate a model's performance. A higher F1-score reflects a better balance between these metrics. Precision measures the accuracy of positive predictions made by the model, while recall (also known as sensitivity or true positive rate) evaluates the model's ability to identify all relevant positive instances. A model with a high F1-score demonstrates good performance in predicting classes in a balanced manner, both for positive and negative cases. Therefore, the best model is determined by considering both accuracy and F1-score, depending on the specific requirements of the classification task.

3. Result and Discussion

The analysis begins by examining the distribution of sentiment labels within the dataset. Figure 1 illustrates a bar chart depicting the frequency distribution of sentiment categories after the initial preprocessing stage, which included removing missing values and duplicate entries. The dataset comprises 6,757 tweets, categorized into two sentiment classes: Positive and Negative.

As shown in the chart, the majority of the tweets are labeled as Positive (5,013 tweets), while the remaining 1,743 tweets are labeled as Negative. This indicates a predominance of positive sentiment within the dataset, suggesting that discussions surrounding the political figure analyzed, Prabowo Subianto, during the specified timeframe leaned more towards favorable opinions or support. Although the dataset exhibits some imbalance in sentiment distribution, the degree of imbalance is not substantial. Therefore, the analysis proceeds without implementing specific techniques to address class imbalance, allowing the models to process the data in its original form.

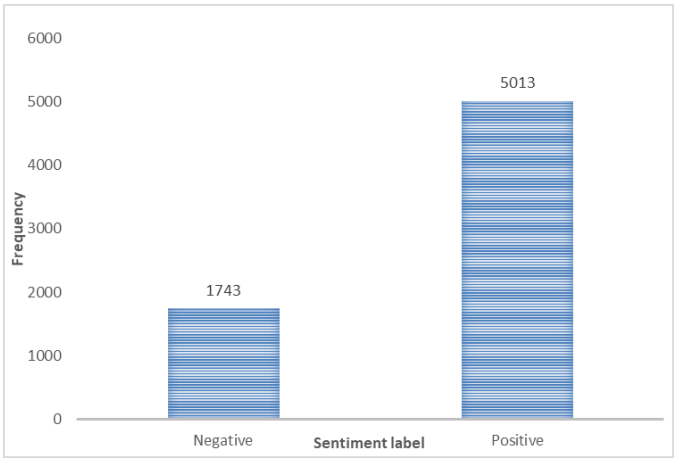

Figure 1: Distribution of dataset based on sentiment category

The result of the training process of sentiment classification using LSTM model is presented in Figure 2. The LSTM model was trained over five epochs using the preprocessed dataset. The training accuracy and loss, along with validation accuracy and loss, were recorded to monitor the model's learning progress. During training, a loss function serves to adjust the model's parameters. It evaluates the discrepancy between the model's predicted outputs and the expected results. The goal of training is to reduce this discrepancy as much as possible (Terven et al, 2024). The detailed results for each epoch are presented below:

1. Training Accuracy and Loss: The training accuracy improved steadily from 0.7281 (72.81%) in Epoch 1 to 0.7498 (74.98%) in Epoch 5, while the training loss decreased from 0.5860 to 0.5648. This trend indicates that the model was able to learn from the data and reduce its prediction error on the training set over successive epochs.
2. Validation Performance: The validation accuracy remained constant at 0.7276 (72.76%) throughout all five epochs, and the validation loss fluctuated slightly between 0.5885 and 0.5858. These results suggest that the model may not have generalized well to the validation set, possibly indicating early stagnation in learning or an inability to capture additional meaningful patterns from the data.

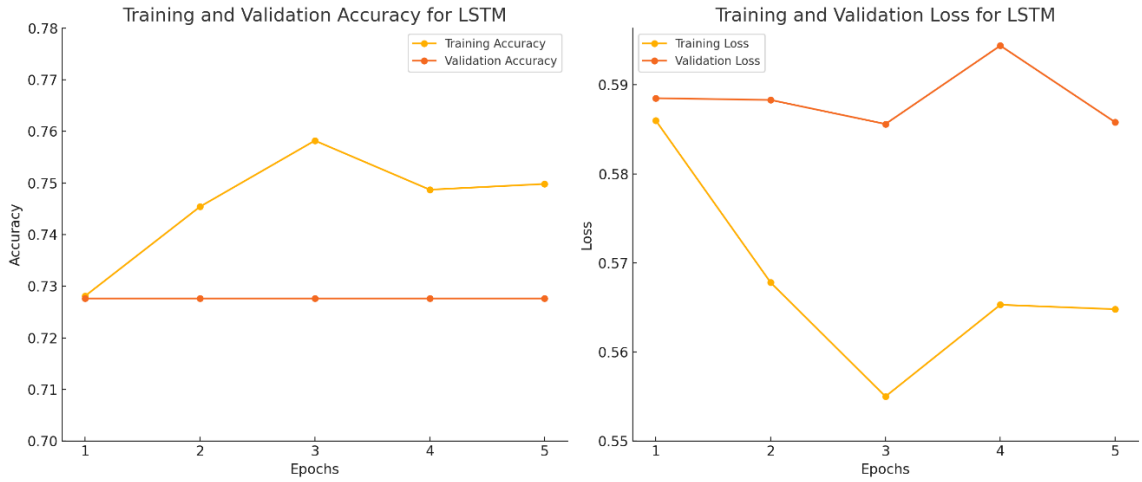

Figure 2: Accuracy and loss plots during the training process for LSTM model

The result of the training process for the BiLSTM model is illustrated in Figure 3. Similar to the LSTM model, the BiLSTM model was trained over five epochs, and the performance metrics were recorded for both the training and validation sets.

1. Training Accuracy and Loss: The BiLSTM model exhibited a notable increase in training accuracy, starting at 73.67% (epoch 1) and reaching 97.27% (epoch 5). Simultaneously, the training loss dropped significantly from 0.5580 to 0.0985, demonstrating the the BiLSTM model was also capable to learn effectively from the training data and minimize prediction errors.

2. Validation Performance: Validation accuracy improved from 0.7550 (75.50%) at epoch 1 to its highest point at 0.8490 (84.90%) at epoch 2 but gradually declined to 0.8275 (82.75%) at epoch 5. Similarly, validation loss initially decreased from 0.4986 to 0.3708 but then increased to 0.4563. These patterns suggest that the model's generalization peaked early, and additional training resulted in overfitting, as reflected in the declining validation metrics.

Overall, the BiLSTM model exhibited better generalization than the LSTM model, as evidenced by its higher validation accuracy and lower validation loss in the initial epochs. This improvement can be attributed to the bidirectional nature of BiLSTM, which processes the input sequence in both forward and backward directions. By doing so, the model captures contextual information from preceding words (forward context) as well as succeeding words (backward context), enabling a more comprehensive understanding of the text data and improving its ability to predict accurately in tasks requiring sequential information. However, fine-tuning the model or implementing regularization techniques could further enhance its validation performance and prevent potential overfitting.

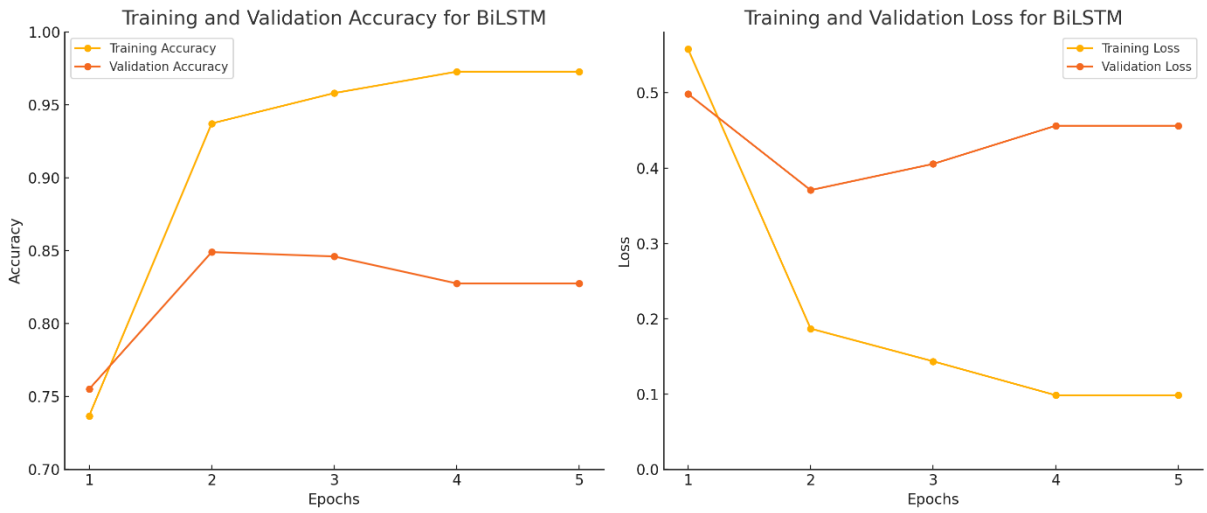

Figure 3: Accuracy and loss plots during the training process for BiLSTM model

Similarly, the training process for the GRU model is summarized in Figure 4. Like the LSTM and BiLSTM models, the GRU model was trained for five epochs, with performance metrics recorded for both the training and validation datasets.

1. Training Accuracy and Loss: The GRU model showed a gradual improvement in training accuracy, increasing from 0.7228 (72.28%) at Epoch 1 to 0.7604 (76.04%) at Epoch 5, while the training loss decreased consistently from 0.5938 to 0.5530. This result also indicates that the GRU model effectively learned from the training data and minimized its prediction error over successive epochs.

2. Validation Performance: The validation accuracy for the GRU model remained constant at 72.76% across all five epochs, similar to the LSTM model. Validation loss fluctuated slightly, starting at 0.5894 at Epoch 1, peaking at 0.6106 at Epoch 3, and then declining to 0.5873 at the final epoch. These results suggest that while the GRU model was able to learn effectively from the training data, it faced challenges in generalizing to the validation set, exhibiting limited improvement in validation performance.

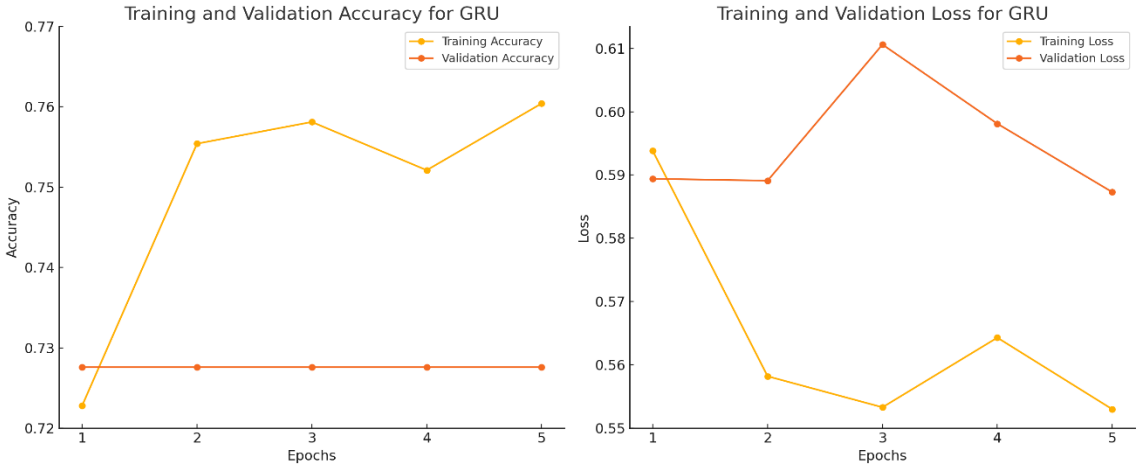

Figure 4: Accuracy and loss plots during the training process for GRU model

Compared to the LSTM and BiLSTM models, the GRU model displayed training accuracy and loss trends similar to LSTM, with relatively stable but stagnant validation accuracy. This pattern reinforces the notion that GRU is comparable to LSTM in performance in certain tasks while requiring fewer parameters and computational resources. However, in this analysis, the GRU model did not achieve the same level of generalization as BiLSTM, further emphasizing the importance of selecting an appropriate model architecture for sentiment analysis tasks.

In this study, the performance of three deep learning models, LSTM, BiLSTM, and GRU, was evaluated for sentiment classification using metrics of test accuracy and F1 score, as presented in Table 1.

The LSTM model achieved a test accuracy of 72.56% and an F1 score of 84.1. These results indicate that while the LSTM model was moderately effective at classifying sentiments, its generalization capability on the test set was limited, reflected in its relatively lower accuracy rather than BiLSTM. The F1 score, which accounts for both precision and recall, suggests the model performed reasonably well at balancing false positives and false negatives.

Table 1: Performance Comparison of Sentiment Classification Models

Model	Test Accuracy (%)	Test F1 Score (%)
LSTM	72.56	84.10
BiLSTM	80.70	86.86
GRU	72.56	84.10

The BiLSTM model outperformed the other models, achieving the highest test accuracy of 80.7% and an F1 score of 86.86. This indicates that the BiLSTM model was better able to capture contextual information from the dataset, leveraging its bidirectional architecture to improve generalization. The higher F1 score further highlights its balanced performance across precision and recall, making it the most effective model for sentiment classification in this study.

The GRU model displayed identical test accuracy (72.56%) and F1 score (84.1) to the LSTM model. While GRU models are known for their computational efficiency due to fewer parameters, this analysis suggests that GRU did not provide significant improvements over LSTM in terms of classification performance. Like LSTM, the GRU model showed limited generalization capability on the test set.

The comparative analysis reveals that the BiLSTM model consistently outperformed both the LSTM and GRU models in terms of test accuracy and F1 score. The BiLSTM model's ability to process information in both forward and backward directions likely contributed to its superior performance, allowing it to capture richer contextual information. This indicates that our results confirm that BiLSTM outperforms LSTM and GRU, further supporting its effectiveness in achieving superior classification performance. Our findings align with those of Ma’aly et al. (2024), which demonstrated that the BiLSTM model achieved the highest accuracy compared to CNN and the Hybrid CNN-BiLSTM model. Rolangon et al. (2023) on their study using twitter dataset about hospital services also obtained that BiLSTM achieve the higher accuracy rather than LSTM and GRU. Giustino & Santosa (2024) on their study about toxic comment classification also concluded that BiLSTM gives the higher performance compared to LSTM, GRU, and Bidirectional Gated Recurrent Unit (BiGRU).

In contrast, the LSTM and GRU models achieved identical performance metrics, suggesting similar capabilities in this specific task. While GRU is computationally less intensive than LSTM, its performance did not provide any notable advantage in terms of accuracy or F1 score in this study.

4. Conclusion and Future Direction

The results of this study indicate that the BiLSTM model outperformed both the LSTM and GRU models in all metrics, achieving a testing accuracy of 80.70% and an F1 score of 86.86%. The higher performance of BiLSTM can be attributed to its ability to capture bidirectional context, which allows for a more nuanced understanding of complex sentiment patterns in social media data. Both LSTM and GRU models demonstrated similar performance, but BiLSTM remains the best choice for sentiment analysis tasks requiring higher accuracy. These findings suggest that BiLSTM is more

effective for analyzing public sentiment regarding political figures, such as Prabowo Subianto, thereby providing valuable insights into public discourse during the 2024 Indonesian Presidential Election.

Future research can explore the use of transformer-based models such as BERT and GPT variant to further improve sentiment classification accuracy. These models are known for their strong contextual understanding, which could enhance the detection of subtle sentiment cues in political discourse. Additionally, experimenting with hyperparameter tuning or incorporating attention mechanisms could further improve the performance of the LSTM, BiLSTM, and GRU models. The potential integration of domain-specific pre-trained models and multilingual datasets could also be valuable for expanding the applicability of the models to diverse linguistic and cultural contexts in sentiment analysis.

References

Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. *Journal of Electrical Systems and Information Technology*, *10*(1), 5. https://doi.org/10.1186/s43067-023-00070-9

Adams, E. J., Ofordi, J., Abdulmumini, A., & Isah, J. M. (n.d.). *The role social media play in generating political awareness, discussion and strategies for better elections*.

Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., & Singh, K. P. (2020). Analysis of Political Sentiment Orientations on Twitter. *Procedia Computer Science*, *167*, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

Bonetti, A., Martínez-Sober, M., Torres, J. C., Vega, J. M., Pellerin, S., & Vila-Francés, J. (2023). Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. *Applied Sciences*, *13*(10), 6038. https://doi.org/10.3390/app13106038

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, *56*(11), 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (arXiv:1406.1078). arXiv. https://doi.org/10.48550/arXiv.1406.1078

Dumlao, J. (2024). *Indonesia Presidential Candidate's Dataset, 2024* [Dataset]. https://www.kaggle.com/datasets/jocelyndumlao/indonesia-presidential-candidates-dataset-2024

Firdaus, A. A., Saputro, J. S., Anwar, M., Adriyanto, F., Maghfiroh, H., Ma'arif, A., Syuhada, F., & Hidayat, R. (n.d.). *Application of Sentiment Analysis as an Innovative Approach to Policy Making: A Review*.

Giustino, J. K., & Santosa, Y. P. (2024). Toxic Comment Classification Comparison between LSTM, BiLSTM, GRU, and BiGRU. *Proxies : Jurnal Informatika*, *7*(2), 115–127. https://doi.org/10.24167/proxies.v7i2.12471

Han, L., Pan, W., & Zhang, H. (2021). Microblog Rumors Detection Based on Bert-GRU. In Q. Liang, W. Wang, J. Mu, X. Liu, Z. Na, & X. Cai (Eds.), *Artificial Intelligence in China* (Vol. 653, pp. 450–457). Springer Singapore. https://doi.org/10.1007/978-981-15-8599-9_52

Hananto, A. L., Nardilasari, A. P., Fauzi, A., Hananto, A., Priyatna, B., & Rahman, Y. (n.d.). Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter. *International Journal of Intelligent Systems and Applications in Engineering*.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Islam, Md. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, Md. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, *57*(3), 62. https://doi.org/10.1007/s10462-023-10651-9

Ma'aly, A. N., Pramesti, D., Fathurahman, A. D., & Fakhrurroja, H. (2024). Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm. *Information*, *15*(11), 705. https://doi.org/10.3390/info15110705

Ragheb, W., Azé, J., Bringay, S., & Servajean, M. (2019). *Attention-based Modeling for Emotion Detection and Classification in Textual Conversations* (arXiv:1906.07020). arXiv. https://doi.org/10.48550/arXiv.1906.07020

Rolangon, A., Weku, A., & Sandag, G. A. (2023). Perbandingan Algoritma LSTM Untuk Analisis Sentimen Pengguna Twitter Terhadap Layanan Rumah Sakit Saat Pandemi Covid-19. *TeIKa*, *13*(01), 31–40. https://doi.org/10.36342/teika.v13i01.3063

Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using deep learning techniques: A comprehensive review. *International Journal of Multimedia Information Retrieval*, *12*(2), 41. https://doi.org/10.1007/s13735-023-00308-2

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Singh, S., & Kumar, P. (2023). Sentiment Analysis of Twitter Data: A Review. *2023 2nd International Conference for Innovation in Technology (INOCON)*, 1–7. https://doi.org/10.1109/INOCON57975.2023.10100998

Suhaeni, C., & Yong, H.-S. (2023). Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences. *Applied Sciences*, *13*(17), 9766. https://doi.org/10.3390/app13179766

Suhaeni, C., & Yong, H.-S. (2024). Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach. *Applied Sciences*, *14*(2), 622. https://doi.org/10.3390/app14020622

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences*, *13*(7), 4550. https://doi.org/10.3390/app13074550

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. (2024). *Loss Functions and Metrics in Deep Learning* (arXiv:2307.02694). arXiv. https://doi.org/10.48550/arXiv.2307.02694

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009