

Enhanced LSTM and BiLSTM-Based Sentiment Analysis of Public Opinion on Ibu Kota Nusantara (IKN)

Kyla Azzahra Kinan ^{1*}, Annisaa Utami ^{2*}, Aminatus Sa'adah ^{3*}

^{*} Informatics Engineering, Telkom University Purwokerto

kylaaazzahra@telkomuniversity.ac.id ¹, annisaa@telkomuniversity.ac.id ², aminatuss@telkomuniversity.ac.id ³

Article Info

Article history:

Received 2026-03-03

Revised 2026-08-15

Accepted 2026-10-03

Keyword:

*BiLSTM,
Deep Learning,
LSTM,
Sentiment Analysis,
Twitter.*

ABSTRACT

The establishment of Ibu Kota Nusantara (IKN) has triggered extensive public discourse on the social media platform X (formerly Twitter), highlighting the importance of systematically evaluating public sentiment toward this national initiative. However, sentiment classification in this context remains challenging due to heterogeneous opinions, informal linguistic usage, and the large-scale, unstructured nature of textual data. This study examines public sentiment related to IKN using deep learning approaches, specifically Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), and compares their performance with a conventional Support Vector Machine (SVM) baseline. Indonesian-language tweets collected from April 2024 to April 2025 via Tweet Harvest were processed through comprehensive preprocessing steps, sentiment labeling using the Indonesia Sentiment Lexicon (InSet), data balancing using SMOTE, and semantic representation via Word2Vec Skip-gram embeddings trained on a domain-specific corpus. The deep learning models were trained for 100 epochs with an embedding dimension of 64, while key hyperparameters such as batch size, neuron units, and dropout rates were systematically optimized. Experimental results show that LSTM achieved a validation accuracy of 0.8934, while BiLSTM performed slightly better with 0.8994, indicating improved contextual learning capability. In contrast, SVM obtained the highest accuracy of 0.91, demonstrating strong performance as a classical machine learning baseline. Both LSTM and BiLSTM models achieved ROC-AUC scores above 0.97, reflecting excellent class separability. Overall, although performance differences among models were relatively marginal, BiLSTM provided more consistent sentiment discrimination across classes, suggesting its effectiveness for analyzing Indonesian social media sentiment regarding the IKN development, while SVM remains a computationally efficient and competitive benchmark model.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. PENDAHULUAN

National strategic policies with broad societal impacts, such as capital relocation, often generate diverse public responses. In the digital era, social media has become a primary channel for individuals to express support, criticism, or concerns regarding major national issues [1]. One widely used platform for this purpose is X (formerly Twitter), which generates short and dynamic textual data. Analyzing public opinion expressed on this platform is essential for

understanding societal perspectives on significant national developments.

The development of Ibu Kota Nusantara (IKN) is a strategic government initiative aimed at relocating Indonesia's administrative center from Jakarta to East Kalimantan, specifically within the Penajam Paser Utara and Kutai Kartanegara regions. The legal foundation for this development is stipulated in Law Number 3 of 2022 on the State Capital, which governs the planning, construction, and relocation processes comprehensively [2]. Beyond serving as a symbol of national identity, the IKN project is intended to

stimulate economic growth in eastern Indonesia, promote equitable development, and reduce structural pressures on Jakarta [3].

From a sustainable development perspective, the IKN project presents considerable challenges, particularly regarding environmental, social, and biodiversity impacts in Kalimantan [4]. The region is recognized for its high biodiversity value, rendering its ecosystems particularly vulnerable to large-scale environmental changes. Consequently, successful implementation of IKN requires consideration not only of economic and physical outcomes but also of ecological and social sustainability.

Sentiment analysis is a computational approach used to classify textual data into polarity categories, such as positive, neutral, or negative [5]. This method enables researchers to discern patterns of public perception and sentiment towards specific issues [6]. In the context of IKN, sentiment analysis provides valuable insights into public support, opposition, or concerns. These insights enable government agencies to respond promptly and refine communication strategies. Furthermore, they assist in anticipating potential challenges and adapting policies to evolving social dynamics.

Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) were selected due to their capability to model complex sequential textual data and capture contextual dependencies. LSTM, an extension of recurrent neural networks (RNNs), effectively addresses the vanishing gradient problem and retains long-term dependencies in sequences [7]. While LSTM processes data in a single direction, BiLSTM enhances its capability by analyzing sequences in both forward and backward directions, allowing for a more comprehensive understanding of contextual information within text [8].

II. MATERIALS AND METHODS

This study analyzes public sentiment regarding the development of Ibu Kota Nusantara (IKN) using data collected from X (formerly Twitter). The research employed a Natural Language Processing (NLP) approach using Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures. The methodological workflow comprised data collection, preprocessing, lexicon-based labeling, word embedding generation, data balancing using SMOTE, data splitting, model training, and evaluation.

Data were collected from Indonesian-language tweets mentioning IKN between 1 April 2024 and 30 April 2025 using the Python-based Tweet Harvest library, which interfaces with the official X API to extract relevant public tweets containing specific keywords such as "IKN," "Pembangunan IKN," and "Ibu Kota Nusantara". A total of 199,826 Indonesian-language tweets related to Ibu Kota Nusantara (IKN) were collected from X during the observation period. The resulting dataset was stored in CSV format, with columns including tweet content, user ID, timestamps, likes, retweets, and other metadata. For analysis,

only the tweet text was used, as it contains the textual information necessary for sentiment classification.

Data preprocessing was conducted to clean and prepare tweet texts prior to sentiment analysis. Several preprocessing stages were applied sequentially, including case folding, cleaning, tokenization, normalization, stopword removal, and stemming. These steps were intended to reduce noise, standardize textual representations, and improve the quality of feature extraction and sentiment classification.

The case folding process transformed all textual data into lowercase format to maintain uniformity in word representation throughout the analysis. Afterward, a cleaning procedure was carried out to eliminate unnecessary components, including hyperlinks, user mentions, hashtags, punctuation symbols, numerical characters, emojis, and other non-alphabetic elements that could potentially introduce noise into the dataset [9]. Once the text had been cleaned, tokenization was performed by segmenting each tweet into separate word-level tokens to facilitate subsequent preprocessing and sentiment classification stages [10].

The normalization phase aimed to standardize informal Indonesian terms, abbreviations, and colloquial expressions commonly found in social media content into their formal linguistic forms. This procedure employed the kamuskatabaku.xlsx dictionary sourced from a public GitHub repository [11], which provides mappings between non-standard and standard Indonesian vocabulary. Through this process, abbreviated expressions such as "klo", "jg", and "ga" were converted into their formal equivalents, namely "kalau", "juga", and "enggak", respectively, to improve textual consistency prior to sentiment analysis.

After the normalization process, stopword removal was carried out to discard frequently occurring Indonesian words with minimal semantic contribution, including conjunctions, prepositions, and other commonly used functional terms. This stage employed the Indonesian stopword corpus provided by the Sastrawi library to retain only informative terms relevant to sentiment analysis. Furthermore, stemming was conducted using the Nazief-Adriani algorithm implemented through the Sastrawi Stemmer library to convert inflected or affixed words into their root forms. For instance, the word "dibubarkan" was reduced to "bubar", whereas "pencitraan" was transformed into "citra" in order to standardize vocabulary representation within the dataset [12].

After completing the preprocessing procedures, sentiment labeling was performed using the InSet Lexicon (Indonesian Sentiment Lexicon) framework. In this approach, each token extracted from the tweets was matched with sentiment-bearing entries contained in the lexicon, where every entry possesses a polarity score ranging from -5 to +5. Positive polarity values indicated favorable sentiment tendencies, whereas negative values represented unfavorable sentiment orientations; a value of zero denoted neutral sentiment. Subsequently, the polarity scores of all tokens within a tweet were aggregated to determine the overall sentiment category of the tweet as positive, negative, or neutral [13].

Although lexicon-based annotation enables efficient large-scale sentiment labeling, the approach remains constrained by the coverage and contextual adequacy of the employed lexicon. In social media discourse, this method often struggles to interpret sarcasm, irony, implicit sentiment, and dynamically evolving informal expressions prevalent in Twitter interactions. Moreover, heuristic labeling mechanisms may introduce annotation noise and polarity distortion when processing slang, multi-word expressions, or context-sensitive linguistic patterns. Consequently, certain tweets may receive sentiment labels that only partially reflect their intended semantic meaning, potentially affecting annotation reliability and introducing latent bias into the dataset [14].

Word representations were generated using Word2Vec with the Skip-gram architecture, configured with a 100-dimensional vector size, a context window of five words, and a minimum word frequency of one. Word2Vec is a neural network-based word embedding method that transforms words into fixed-dimensional numerical vectors, capturing semantic meaning and contextual relationships between words within a corpus [15]. The Skip-gram model aims to predict surrounding context words given a target word, thereby learning word vectors that reflect how words co-occur in sentences or documents. In this study, the Word2Vec model was trained from scratch using the entire dataset, rather than employing pre-trained embeddings, in order to better capture domain-specific linguistic characteristics. This approach is supported by the relatively large size of the dataset, which is sufficient for learning meaningful word representations, as also reported in prior studies where domain-specific corpora benefit from locally trained embeddings. Furthermore, training from scratch allows the model to better adapt to informal language patterns, slang, and contextual nuances commonly found in social media text, which may not be well represented in general-purpose pre-trained embeddings. Document vectors for individual tweets were obtained by averaging the vectors of all words in each tweet, producing dense, semantically meaningful representations suitable for subsequent sentiment analysis and machine learning tasks [16].

For modeling, the dataset was split into 80% training and 20% testing. This proportion was selected to ensure a sufficient amount of data for model learning while preserving an adequate portion for unbiased evaluation on unseen data. The large dataset size further supports the reliability of this split in representing the overall data distribution. This study employed a train-test split approach and did not implement k-fold cross validation. Additionally, 20% of the training data was further allocated as validation data during training using the validation split mechanism to monitor model generalization performance.

Two types of deep neural network architectures, Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM), were used to build sentiment classification models. LSTM is a variant of Recurrent Neural

Networks (RNNs) designed to overcome the vanishing gradient problem, allowing the model to capture long-term dependencies in sequential data [17]. The LSTM architecture employs gated mechanisms (input, forget, and output gates) to regulate information flow and mitigate the vanishing gradient problem. Through these gated mechanisms, the model selectively retains or discards information, enabling effective learning of long-range dependencies in textual sequences [18]. BiLSTM extends this architecture by processing sequences in both forward and backward directions, thereby capturing bidirectional contextual dependencies within textual data [19]. Multiple configurations of LSTM and BiLSTM were trained using the Adam optimizer with a learning rate of 0.001, while varying the number of epochs, batch sizes, and neurons to identify the optimal architecture based on validation performance.

The models were evaluated using confusion matrices alongside accuracy, precision, recall, F1-score, and ROC-AUC with a macro-average One-vs-Rest approach to assess their ability to distinguish sentiment classes effectively. A confusion matrix provides a detailed representation of prediction outcomes by comparing actual and predicted class labels, showing true positives, true negatives, false positives, and false negatives, which helps identify where misclassifications occur and assess performance across each class [20].

ROC-AUC evaluates the model's discriminative capability by measuring the trade-off between the true positive rate and false positive rate across classification thresholds. Higher AUC values indicate stronger class separability, making this metric suitable for multi-class sentiment classification using a One-vs-Rest scheme. The combination of confusion matrix analysis and ROC-AUC provides both class-level interpretability and overall discrimination assessment, enabling a comprehensive evaluation of model performance [21].

III. RESULTS AND DISCUSSION

A. Data and Label

A total of 199,826 Indonesian-language tweets related to Ibu Kota Nusantara (IKN) were collected from X during the observation period. After preprocessing, sentiment labeling was conducted using the Indonesia Sentiment Lexicon (InSet) by aggregating token-level polarity scores

TABEL I
LABELLING RESULTS

No	Text	Polarity Score	Label
1	[klaim, investor, ikn, antri, nol]	-3	Negative
2	[asik, ikn, barter, pemirsah]	0	Neutral
3	[kongres, solo, ikn]	2	Positive

Table 1 presents sample results of the lexicon-based sentiment labeling process. The Text column contains tokenized and preprocessed words after stopword removal

and stemming. Polarity scores were calculated by aggregating the sentiment weights of words based on the Indonesian Sentiment Lexicon (InSet).

Tweets with negative aggregate scores were classified as Negative (e.g., a score of -3), positive aggregate scores were classified as Positive (e.g., a score of 2), and zero scores were categorized as Neutral. This scoring mechanism provides a systematic basis for assigning sentiment labels prior to model training.

B. Visualization

Visualizations were generated to present the most frequently occurring words in the dataset and to illustrate the distribution of sentiment labels (Negative, Positive, and Neutral). A total of 199,826 tweets were analyzed to capture the overall sentiment landscape of discussions related to Ibu Kota Nusantara (IKN).

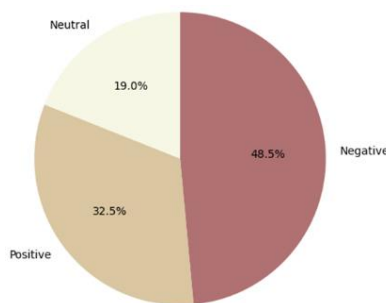


Figure 1. Distribution Label

Figure 1 displays the distribution of sentiment labels across the dataset. Negative sentiment dominates with 96,967 tweets (48.5%), followed by Positive sentiment with 64,927 tweets (32.5%) and Neutral sentiment with 37,932 tweets (19%). These findings indicate a predominance of negative sentiment in public discourse regarding IKN.

C. Word Embedding

The study employed the Skip-gram architecture of Word2Vec as the word embedding technique to convert textual data into dense numerical vector representations. This representation captures semantic relationships by placing contextually similar words in close proximity within the vector space, thereby enhancing the effectiveness of subsequent classification modeling.

D. SMOTE

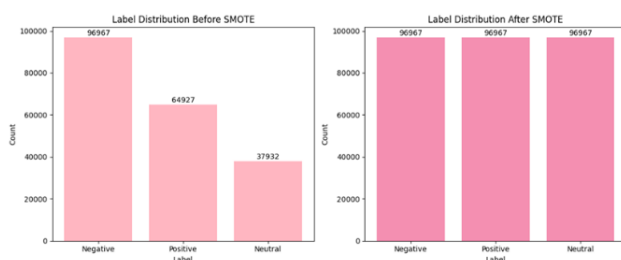


Figure 2. Label Distribution Before and After SMOTE

Figure 2 presents the class distribution before and after the application of SMOTE. Initially, the dataset exhibited class imbalance, with the negative class (96,967 samples) substantially exceeding the positive (64,927) and neutral (37,932) classes. Class imbalance can introduce bias toward the majority class and reduce the model's ability to accurately learn minority class patterns.

To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic instances for the positive and neutral classes. After oversampling, each sentiment category contained 96,967 samples, resulting in a balanced training distribution. This resampling process is expected to support more stable and equitable learning across all sentiment classes.

After applying SMOTE, the dataset was divided into training and testing sets using an 80:20 split. From the total 290,901 samples, 232,720 were allocated for training and 58,181 for testing. This proportion follows common practice in supervised learning to balance training adequacy and evaluation reliability.

E. Model

The LSTM and BiLSTM models were trained using the Adam optimizer with a fixed learning rate of 0.001 due to its proven effectiveness in ensuring stable convergence during training. [All experimental scenarios were conducted for 100 epochs with an embedding dimension of 64, with the hyperparameter configurations presented in Table 2. Variations in batch size and neuron units were introduced to examine their impact on model performance. Therefore, the hyperparameter tuning process focused on batch size and neuron configurations to analyze their influence on model performance while maintaining training stability.

TABEL II
HYPERPARAMETER TUNING

Scen.	Epoch	Batch Size	Embedding Size	Neuron
1	100	32	64	32
2	100	64	64	64
3	100	32	64	64
4	100	64	64	128
5	100	32	64	128

To mitigate overfitting, dropout rates of 0.3 and 0.4 were applied, and a validation split of 20% was used during training to monitor model performance on unseen data at each epoch. The sparse categorical crossentropy loss function was employed, as it is suitable for multi-class classification problems with integer labels. Early stopping was not applied in this study; however, dropout regularization, validation monitoring, and a learning rate of 0.001 contributed to stable convergence and reduced the risk of overfitting. The optimal model was selected based on validation accuracy and validation loss rather than training performance to ensure better generalization on unseen data.

TABEL III
LSTM RESULTS WITH DROPOUT 0.3

Scen.	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	0.9196	0.8910	0.2274	0.3244
2	0.9538	0.8928	0.1385	0.4012
3	0.9618	0.8869	0.1147	0.4315
4	0.9917	0.8912	0.0276	0.6210
5	0.9819	0.8907	0.0554	0.4776

As shown in Table 3, Scenario 1 provided the best trade-off between training and validation performance among all LSTM configurations with a dropout rate of 0.3. Although several scenarios achieved higher training accuracy, their validation accuracy did not improve proportionally and was accompanied by increased validation loss. Scenario 1 maintained relatively stable performance between training accuracy (0.9196) and validation accuracy (0.8910), while also producing a lower validation loss compared to more complex configurations. These results indicated better generalization capability, whereas configurations with larger neuron sizes showed stronger tendencies toward overfitting.

TABEL IV
BiLSTM RESULTS WITH DROPOUT 0.3

Scen.	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	0.9302	0.8859	0.1959	0.3660
2	0.9753	0.8865	0.0718	0.5501
3	0.9769	0.8896	0.0679	0.5282
4	0.9934	0.8946	0.0209	0.6993
5	0.9874	0.8990	0.0379	0.5685

As presented in Table 4, Scenario 5 achieved the highest validation accuracy (0.8990) among the evaluated BiLSTM configurations with a dropout rate of 0.3. Although several scenarios produced extremely high training accuracy, they were associated with substantially higher validation loss, particularly in configurations with larger neuron sizes. Scenario 5 demonstrated more stable validation performance and better generalization capability compared to other BiLSTM configurations.

TABEL V
LSTM RESULTS WITH DROPOUT 0.4

Scen.	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	0.9147	0.8934	0.2368	0.3051
2	0.9586	0.8941	0.1269	0.4035
3	0.9605	0.8849	0.1183	0.4478
4	0.9864	0.8918	0.0481	0.5447
5	0.9907	0.8916	0.0305	0.6159

As presented in Table 5, Scenario 1 demonstrated the most balanced performance among the LSTM configurations with a dropout rate of 0.4. Although its training accuracy (0.9147) was lower than more complex scenarios, it achieved the lowest validation loss (0.3051) and maintained a relatively

small gap between training and validation accuracy (0.9147 vs 0.8934). In contrast, configurations with higher neuron counts yielded substantially higher training accuracy but were accompanied by increased validation loss, indicating potential overfitting. Therefore, Scenario 1 was considered the most suitable configuration under the 0.4 dropout setting.

TABEL VI
BiLSTM RESULTS WITH DROPOUT 0.4

Scen.	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	0.9317	0.8815	0.1938	0.3885
2	0.9721	0.8829	0.0801	0.5526
3	0.9729	0.8830	0.0793	0.5207
4	0.9913	0.8994	0.0252	0.6369
5	0.9856	0.8945	0.0427	0.5690

As shown in Table 6, Scenario 4 achieved the highest validation accuracy (0.8994) among the BiLSTM configurations with a dropout rate of 0.4. Despite exhibiting a relatively large gap between training and validation accuracy, this configuration outperformed other scenarios in terms of validation accuracy, which served as the primary selection criterion. Although its validation loss was higher than some alternative configurations, the superior validation accuracy suggests that Scenario 4 provided better predictive performance on unseen data compared to other settings.

Based on the comparative analysis of training accuracy, validation accuracy, training loss, and validation loss, configurations using a dropout rate of 0.4 demonstrated more stable validation performance compared to those using a dropout rate of 0.3. For the LSTM model, Scenario 1 with dropout 0.4 achieved reliable performance (training accuracy = 0.9147; validation accuracy = 0.8934) and produced the lowest validation loss (0.3051) among all evaluated configurations. In contrast, several configurations with dropout 0.3 achieved higher training accuracy but were associated with increased validation loss, suggesting reduced generalization capability.

For the BiLSTM model, Scenario 4 with dropout 0.4 achieved the highest validation accuracy (0.8994). Although this configuration showed a noticeable gap between training and validation accuracy, its validation performance exceeded that of the corresponding configurations with dropout 0.3. Therefore, a dropout rate of 0.4 was considered more effective for improving model generalization in both LSTM and BiLSTM architectures prior to final evaluation.

In the detailed class-level evaluation, the LSTM model achieved its strongest performance on the negative class (precision = 0.91; recall = 0.90), while slightly lower F1-scores were observed for the neutral and positive classes (0.88–0.89). These results indicated that the LSTM model was more effective in identifying negative sentiment patterns compared to the other sentiment categories.

Conversely, the BiLSTM model demonstrated more consistent F1-scores across all sentiment classes (0.89–0.91), particularly improving performance on the neutral class.

These findings suggested that bidirectional sequence modeling enabled the model to capture contextual information more effectively, resulting in more balanced classification performance across sentiment categories, even though the overall accuracy of both models remained relatively comparable.

F. Confusion Matrix

To provide a more detailed evaluation of classification performance, confusion matrix analysis was conducted using the best-performing configurations of the LSTM and BiLSTM models. The selected models were LSTM with dropout 0.4 (Scenario 1) and BiLSTM with dropout 0.4 (Scenario 4), which previously demonstrated the most stable and balanced performance across evaluation metrics.

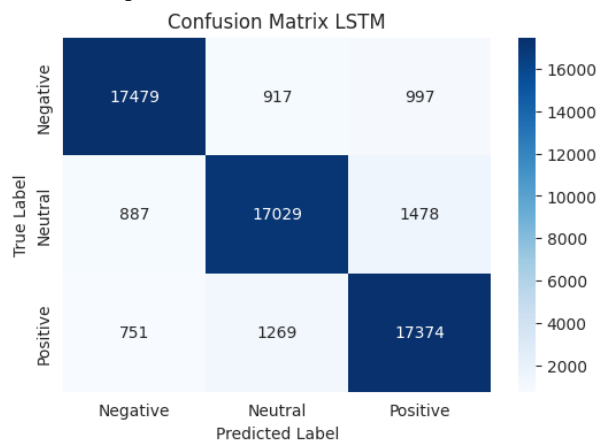


Figure 3. Confusion Matrix of LSTM Model (Scenario 1, Dropout 0.4)

Figure 3 presents the confusion matrix of the best-performing LSTM model with dropout 0.4 (Scenario 1). Overall, the model demonstrated balanced classification performance across all sentiment categories, as reflected by the relatively even distribution of correct predictions along the diagonal of the matrix. The negative class produced the highest number of correct predictions, while several misclassifications were observed between the neutral and positive classes. Specifically, 1,478 neutral tweets were classified as positive, whereas 1,269 positive tweets were predicted as neutral. This pattern suggests that neutral and positive sentiments may share similar linguistic characteristics within the tweet dataset, making them more difficult to distinguish. Nevertheless, the overall prediction distribution remained relatively balanced across all classes, indicating stable sentiment classification performance without significant bias toward any particular category.

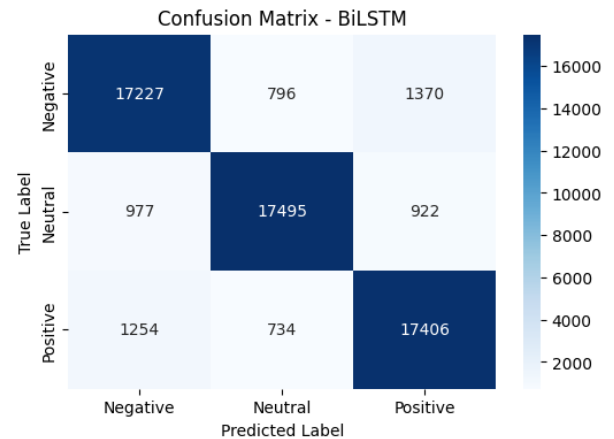


Figure 4. Confusion Matrix of BiLSTM Model (Scenario 4, Dropout 0.4)

Figure 4 presents the confusion matrix of the best-performing BiLSTM model with dropout 0.4 (Scenario 4). Overall, the model demonstrated balanced classification performance across all sentiment categories, as reflected by the relatively even distribution of correct predictions along the diagonal of the matrix. The neutral class achieved the highest number of correct predictions, indicating that the BiLSTM architecture was able to capture contextual information more effectively, particularly for neutral sentiment classification. Several misclassifications were still observed between the negative and positive classes, where 1,370 negative tweets were classified as positive and 1,254 positive tweets were predicted as negative. Nevertheless, the overall distribution of predictions remained relatively balanced across all sentiment categories, suggesting that the bidirectional sequence modeling mechanism contributed to more stable and consistent sentiment classification performance without significant bias toward any particular class.

G. ROC-AUC

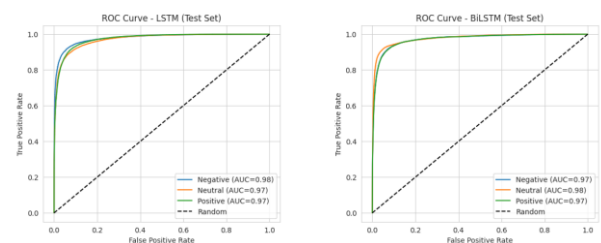


Figure 5. ROC-AUC Results of LSTM and BiLSTM with Dropout 0.3

Figure 5 presents the ROC-AUC performance of the LSTM and BiLSTM models with a dropout rate of 0.3. Both models achieved high AUC values across all sentiment classes. The LSTM model obtained AUC scores of 0.974 for the negative class and 0.970 for both the neutral and positive classes. Similarly, the BiLSTM model achieved an AUC score of 0.973 for the negative class and 0.970 for the remaining classes. These findings indicate that both architectures

demonstrated strong discriminative capability, with only marginal performance differences under the 0.3 dropout configuration.

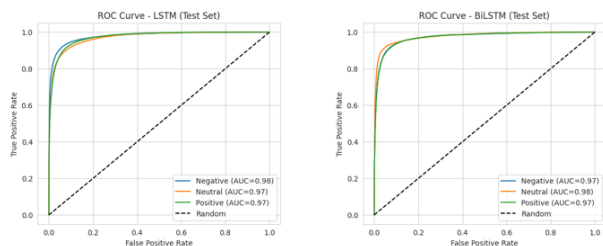


Figure 6. ROC-AUC Results of LSTM and BiLSTM with Dropout 0.4

Figure 6 illustrates the ROC-AUC results for the 0.4 dropout configuration. The LSTM model achieved AUC values of 0.975 for the negative class and 0.970 for both the neutral and positive classes. Meanwhile, the BiLSTM model produced comparable performance, obtaining an AUC score of 0.972 for the negative class and 0.970 for the other sentiment classes. Overall, both dropout configurations resulted in consistently high ROC-AUC values, suggesting that the models maintained stable class discrimination performance across different dropout settings.

Overall, the BiLSTM model with a dropout rate of 0.4 demonstrated the most robust and balanced performance across evaluation metrics, including validation accuracy, class-level F1-scores, and ROC-AUC values. Although the overall performance differences between LSTM and BiLSTM were relatively small, the bidirectional architecture provided more balanced sentiment discrimination across all classes. Therefore, the BiLSTM model with dropout 0.4 was selected as the optimal configuration for sentiment classification of IKN-related tweets.

H. Baseline Comparison Using SVM

To further evaluate the effectiveness of the proposed deep learning architectures, an additional experiment was conducted using Support Vector Machine (SVM) as a baseline classifier. SVM was selected because it is widely recognized as one of the most effective traditional machine learning algorithms for text classification and sentiment analysis, particularly in handling high-dimensional textual feature spaces.

The SVM experiment followed the same initial pipeline as the deep learning models, including data collection, preprocessing, lexicon-based labeling, class balancing using SMOTE, and train-test splitting. However, unlike the LSTM and BiLSTM models that utilized Word2Vec embeddings for sequential representation learning, the SVM model employed TF-IDF feature extraction to transform textual data into numerical vectors suitable for conventional machine learning classification. TF-IDF was configured using unigram and bigram features ($ngram_range = (1,2)$) to capture both individual words and short phrase patterns within tweets. The maximum number of features was limited to 5000 to reduce dimensionality and improve computational efficiency. In

addition, terms appearing in fewer than two documents ($min_df = 2$) and terms occurring in more than 90% of documents ($max_df = 0.9$) were excluded to minimize noise and overly common terms.

The SVM model employed a linear kernel with $C = 1$ due to its computational efficiency and effectiveness for sparse textual representations. The regularization parameter C was selected to balance margin maximization and classification error minimization. Furthermore, the model was trained using the same balanced dataset and identical train-test split configuration (80:20) as the LSTM and BiLSTM models to ensure a fair performance comparison across all approaches.

The experimental results demonstrated that SVM achieved competitive performance, obtaining an accuracy of 0.91 along with balanced precision, recall, and F1-score values across all sentiment classes. The confusion matrix showed that the SVM model classified the negative class most effectively, with 18,001 correctly predicted instances. Several misclassifications were observed between the neutral and positive classes, where 1,478 positive tweets were predicted as neutral and 946 neutral tweets were classified as positive. This pattern suggests that neutral and positive sentiments may share similar linguistic characteristics within the tweet dataset. Nevertheless, the overall prediction distribution remained relatively balanced across all sentiment categories, indicating that the SVM model maintained stable classification performance without significant bias toward any particular class.

Although the performance differences among SVM, LSTM, and BiLSTM were relatively small, the BiLSTM architecture demonstrated advantages in capturing bidirectional contextual dependencies within tweet sequences. In contrast, SVM relied on fixed feature representations and did not explicitly model sequential contextual information. Nevertheless, the strong performance achieved by SVM indicated that traditional machine learning approaches remain competitive as baseline methods for Indonesian sentiment analysis tasks.

I. Computational Complexity

From a computational perspective, deep learning architectures generally require greater computational resources than traditional machine learning approaches such as SVM. Among the evaluated architectures, BiLSTM introduced higher computational overhead than standard LSTM because sequence information was processed in both forward and backward directions. Consequently, BiLSTM required longer training time and higher memory consumption, whereas LSTM was relatively more computationally efficient while still maintaining competitive classification performance.

J. Practical Implications

The findings of this study also provide practical implications for policymakers, particularly in the context of IKN development. The dominance of negative sentiment

indicated that public concerns toward the IKN project remain prevalent and should be addressed through more transparent communication and inclusive policy strategies. By leveraging sentiment analysis results, the government can identify key public issues and monitor opinion trends in real time. These insights may support the development of more targeted communication strategies aimed at improving public trust and engagement. Furthermore, the findings may contribute to evidence-based decision-making processes to ensure that the implementation of the IKN project aligns with public expectations and broader socio-environmental considerations.

Nevertheless, this study still has several limitations. The developed models were evaluated only on Indonesian tweets related to the IKN issue; therefore, the external validity and generalization capability of the models on other datasets, domains, or social media platforms remain limited. In addition, this study primarily focused on predictive performance and did not extensively explore model interpretability. Although BiLSTM demonstrated slightly more balanced classification performance compared to LSTM, this study did not conduct statistical significance testing to formally verify whether the observed performance differences between the models were statistically significant. Therefore, the reported performance improvements should be interpreted as empirical observations within the experimental setting. Since sentiment analysis results may support public policy evaluation and decision-making processes, future research may incorporate explainable artificial intelligence (XAI) approaches to provide more transparent and interpretable sentiment classification results. Future studies may also employ statistical hypothesis testing methods, such as McNemar test or paired significance analysis, and evaluate the proposed approaches on more diverse datasets and broader sentiment analysis contexts to further assess model robustness and generalizability.

Nevertheless, this study still has several limitations. The developed models were evaluated only on Indonesian tweets related to the IKN issue; therefore, the external validity and generalization capability of the models on other datasets, domains, or social media platforms remain limited. In addition, this study primarily focused on predictive performance and did not extensively explore model interpretability. Since sentiment analysis results may support public policy evaluation and decision-making processes, future research may incorporate explainable artificial intelligence (XAI) approaches to provide more transparent and interpretable sentiment classification results. Future studies may also evaluate the proposed approaches on more diverse datasets and broader sentiment analysis contexts to further assess model robustness and generalizability.

IV. CONCLUSION

Based on the experimental results, the LSTM, BiLSTM, and SVM models demonstrated strong performance in the sentiment classification of Indonesian tweets related to IKN. For the deep learning architectures, a dropout rate of 0.4 provided more stable validation performance and effectively reduced overfitting. Among the evaluated LSTM configurations, Scenario 1 demonstrated reliable generalization capability through balanced training and validation performance. Meanwhile, the BiLSTM model in Scenario 4 achieved the highest validation accuracy and exhibited more balanced contextual representation learning through bidirectional sequence processing.

In addition, the SVM baseline model also demonstrated strong classification performance, achieving an accuracy of 0.91 with balanced precision, recall, and F1-score values across all sentiment classes. These findings indicate that traditional machine learning approaches remain effective for Indonesian sentiment analysis tasks, particularly when combined with appropriate preprocessing and TF-IDF feature extraction techniques.

Furthermore, both LSTM and BiLSTM achieved excellent discriminative capability, with ROC-AUC values exceeding 0.97 across all sentiment categories. Although the overall performance differences among the evaluated models were relatively small, BiLSTM demonstrated the most robust and balanced performance across evaluation metrics. Therefore, BiLSTM was identified as the most effective model for capturing contextual sentiment patterns in Indonesian social media text related to IKN, while SVM remained a computationally efficient and competitive baseline approach.

REFERENCES

- [1] M. R. Amin and A. N. Salma, "Analisis Sentimen Opini Publik Mengenai Tragedi Kanjuruhan Pada Media Sosial Twitter Menggunakan Brand24," *eProceedings of Management*, vol. 11, no. 6, pp. 7283–7287, 2024.
- [2] E. Benia and G. Nabilah, "Politik hukum dalam proses pemindahan ibu kota negara melalui pembentukan Undang-Undang Ibu Kota Negara (UU IKN)," *Jurnal Hukum Lex Generalis*, vol. 3, no. 10, pp. 806–825, 2022.
- [3] C. A. Prianggodo, "Pembangunan IKN sebagai mercusuar strategis (kajian kritis dalam perspektif sustainable development)," *Jurnal Pemerintahan dan Politik*, vol. 8, no. 1, 2023.
- [4] T. R. Dewi, N. A. Ramadhani, B. O. Shakira, D. K. N. Jamal, M. Aisyah, I. F. Hidayat, and D. Angela, "Analisis Deep Ecology Arne Naess Terhadap Pembangunan Ibu Kota Nusantara Untuk Melindungi Ekosistem Dan Pembangunan Berkelanjutan," *Journal of Government: Manajemen Pemerintahan dan Otonomi Daerah*, vol. 9, no. 2, pp. 1–19, 2024.
- [5] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 18, no. 5, pp. 42–61, 2023.
- [6] D. P. Jennifer, N. C. Ningrum, S. Pinasty, S. M. N. Edma, W. N. Andini, and E. Widodo, "Analisis Sentimen Pada Pengguna Tiktok Menggunakan Metode Random Forest (Studi Kasus: Jessica-Mirna)," *INNOVATIVE: Journal of Social Science Research*, vol. 4, no. 3, pp. 14477–14489, 2024.

- [7] M. D. A. Nasir and S. Hidayat, "Analisis Sentimen Ulasan Film Menggunakan Metode BiLSTM," *Jurnal Informatika dan Teknologi Komputer (J-ICOM)*, vol. 5, no. 2, pp. 126–132, 2024.
- [8] A. N. Romadhan, E. Utami, and A. D. Hartanto, "Analisis Sentimen Opini Publik Menggunakan Metode BiLSTM Pada Media Sosial Twitter," *Seminar Nasional Teknologi Informasi dan Matematika (SEMIOTIKA)*, vol. 2, no. 1, pp. 22–33, 2023.
- [9] F. Abbas and A. Taeihagh, "A multi-level fusion-based framework for multimodal fake news classification using semantic feature extraction," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 9, pp. 6531–6560, 2025, doi: 10.1007/s13042-025-02633-w.
- [10] A. L. I. Erkan and T. Güngör, "Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification," *IEEE Access*, vol. 11, pp. 134951–134968, 2023.
- [11] A. Hendra, "Twitter Sentiment Analysis Using InSet and Random Forest," GitHub, [Online]. Available: <https://github.com/agushendra7/twitter-sentiment-analysis-using-inset-and-random-forest/tree/main/dataset>
- [12] S. Saifullah, R. Drezewski, F. A. Dwiyanto, A. S. Aribowo, Y. Fauziah, and N. H. Cahyana, "Automated Text Annotation Using a Semi-Supervised Approach with Meta Vectorizer and Machine Learning Algorithms for Hate Speech Detection," *Applied Sciences*, vol. 14, no. 3, pp. 1–19, 2024.
- [13] M. F. N. Fathoni, E. Y. Puspaningrum, and A. N. Sihananto, "Perbandingan performa labeling lexicon InSet dan VADER pada analisis sentimen Rohingya di aplikasi X dengan SVM," *Modem: Jurnal Informatika dan Sains Teknologi*, vol. 1, no. 3, pp. 62–76, 2024.
- [14] M. Sykora, S. Elayan, and T. W. Jackson, "A qualitative analysis of sarcasm, irony and related #hashtags on Twitter," *Big Data & Society*, vol. 7, no. 2, pp. 1–15, 2020, doi: 10.1177/2053951720972735.
- [15] A. Somani, A. Horsch, and D. K. Prasad, *Interpretability in Deep Learning*, 1st ed. Cham, Switzerland: Springer, 2023.
- [16] K. Denecke, *Sentiment Analysis in the Medical Domain*, 1st ed. Cham, Switzerland: Springer, 2023.
- [17] A. Rolangon, A. Weku, and G. A. Sandag, "Perbandingan Algoritma LSTM Untuk Analisis dan Pengguna Twitter Terhadap Layanan Rumah Sakit Saat Pandemi Covid-19," *Jurnal Teknologi Informasi dan Komunikasi*, vol. 13, no. 1, pp. 31–40, 2023.
- [18] S. S. Nurashila, F. Hamami, and T. F. Kusumasari, "Perbandingan kinerja algoritma recurrent neural network (RNN) dan long short-term memory (LSTM): studi kasus prediksi kemacetan lalu lintas jaringan PT XYZ," *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI)*, vol. 8, no. 3, pp. 864–877, 2023.
- [19] A. Muzakir, K. Adi, and R. Kusumaningrum, *Penerapan Konsep Machine Learning & Deep Learning*. Semarang: UNDIP Press, 2024.
- [20] F. R. Valerian, M. Syarieff, and D. A. Fatah, "Klasifikasi tingkat obesitas menggunakan metode GBM dan confusion matrix," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 2242–2249, 2025.
- [21] H. Cam, A. V. Cam, U. Demirel, and S. Ahmed, "Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers," *Heliyon*, vol. 10, no. 1, pp. 1–15, 2024.