Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 3, August, Pp. 629-640

Comparison of Word2Vec and GloVe performance in Bi-LSTM models for Indonesian news classification

Muhammad Faris Wafda[1], Husni[1], Ika Oktavia Suzanti[1], Firdaus Solihin[1], Mula'ab[1], Army Justitia[2]
University of Trunodjoyo Madura, Madura, Indonesia[1]
National Cheng Kung University, Tainan, Taiwan[2]

Article Info
Keywords:
Text Classification, Bi-LSTM, Word2Vec, GloVe, Indonesian News

Article history:
Received: December 03, 2025
Accepted: May 06, 2026
Published: August 01, 2026

Cite:
M. F. Wafda, Husni, Ika Oktavia Suzanti, F. Solihin, Mula'ab, and Army Justitia, "Comparison of Word2Vec and GloVe Performance in Bi-LSTM Models for Indonesian News Classification", *KINETIK*, vol. 11, no. 3, Aug. 2026.
https://doi.org/10.22219/kinetik.v11i3.2608

*Corresponding author.
Muhammad Faris Wafda
E-mail address:
220411100039@student.trunojoyo.ac.id

Abstract
The explosion in the volume of textual data from digital news presents challenges in classifying content automatically and efficiently. For the task of classifying Indonesian-language news, this study aims to compare the performance of several word embeddings specifically Word2Vec using CBOW and Skip-Gram architectures and GloVe when applied to a Bidirectional Long Short-Term Memory (Bi-LSTM) model. This study uses a dataset consisting of 6,715 news articles from the Indonesian news portal that have undergone pre-processing, divided into five categories. The model was trained using 80% of the training data with K-Fold Cross Validation (K=5), while the remaining 20% of the data was used for testing. The experimental findings indicate that the Bi-LSTM model, when combined with CBOW embedding, yielded the best performance, achieving 95.16% accuracy and a 95.15% F1-Score. The Skip-Gram model followed with solid performance, achieving an accuracy of 93.30% and the fastest computation time. Conversely, the model that used pre-trained GloVe embedding delivered the poorest performance, achieving 88.98% accuracy. This result suggests that training embeddings on a specific domain is more effective at capturing local context. The conclusion of this study confirms that selecting a word embedding method specifically trained on local datasets is also an important step in achieving optimal accuracy in Indonesian news text classification.

1. Introduction

Digital news that spreads rapidly through various platforms, such as official news portals and social media, has led to an explosion in the volume of textual data [1]. The Reuters Institute Digital News Report 2025 reports that the use of social media as a source of news has surpassed television in the United States, with 54% of respondents accessing news through social media, compared to 50% through television. This is also reflected in Indonesia, where around 57% of Indonesians rely on social media for news, 40% of whom use social media as their main source of news [2]. This condition poses a challenge in automatically organizing and classifying news. Currently, news categorization on many platforms is still very general. As a result, readers who are looking for specific information must manually filter news from these general categories to find more detailed topics, which is an inefficient and time-consuming process. Therefore, accurate and detailed automatic news classification is important to improve accessibility and information management [3].

Natural language processing offers various solutions in text data processing, one of which is news classification. News classification aims to categorize news text documents into predefined classes based on their content [4]. With an automatic classification system, news portals can present more structured content, making it easier for readers to find information.

In text classification systems, model performance is heavily influenced by how text is represented numerically. This process is known as feature extraction or word embedding [5]. n deep learning, various word embedding methods exist that produce word vector representations, notably Word2Vec and GloVe [6]. ord2Vec learns word representations based on the surrounding context using two main architectures: Continuous Bag-of-Words (CBOW), which predicts the target word based on context, and Skip-gram, which predicts context based on the target word [7]. On the other hand, GloVe relies on a global word co-occurrence matrix to capture global statistics from the entire dataset [8]. Word2Vec is superior in capturing local semantic relationships, whereas GloVe is faster in training and more effective at capturing word analogies based on global statistics [9].

Once words are represented as numerical vectors, deep learning approaches demonstrate superior performance in automatically learning relevant features for text classification tasks [10]. In sequential data processing, Long Short-Term Memory (LSTM) based architectures were developed to achieve a more optimal understanding of text [11]. As an improvement to this base architecture, the Bidirectional LSTM (Bi-LSTM) model is applied to process sequence data

from two directions forward and backward. This dual approach allows Bi-LSTM to capture information more comprehensively and understand broader contextual relationships within the entire text [12].

The fundamental difference in word vector representation between Word2Vec, which learns local context, and GloVe, which learns global statistics, causes the two to have different ways of capturing relationships between words. Differences in learning relationships between words in the domain of Indonesian-language news, which has a distinctive sentence structure, become a problem. Furthermore, high vocabulary overlap between categories like Travel, Money, and Education blurs the boundaries for accurate classification. Additionally, lengthy introductory narratives often obscure the core topic, making precise feature extraction difficult. Currently, there is still a research gap comparing the performance of embedding methods to provide feature representations that can deliver the best accuracy for Bi-LSTM models in Indonesian text classification. To address these challenges, this study proposes a novel approach by restricting the input sequence to 100 tokens to test model robustness, demonstrating that an independently trained domain-specific embedding effectively overcomes the global context bias of pre-trained models. Therefore, this study aims to evaluate the effectiveness of the Bi-LSTM model combined with Word2Vec and GloVe embeddings for Indonesian news classification, so that it can provide clear recommendations for the development of automatic text classification systems in the future.

2. Research Method

This research was conducted through several systematic stages to ensure that the research objectives were achieved. These stages are illustrated in Figure 1.

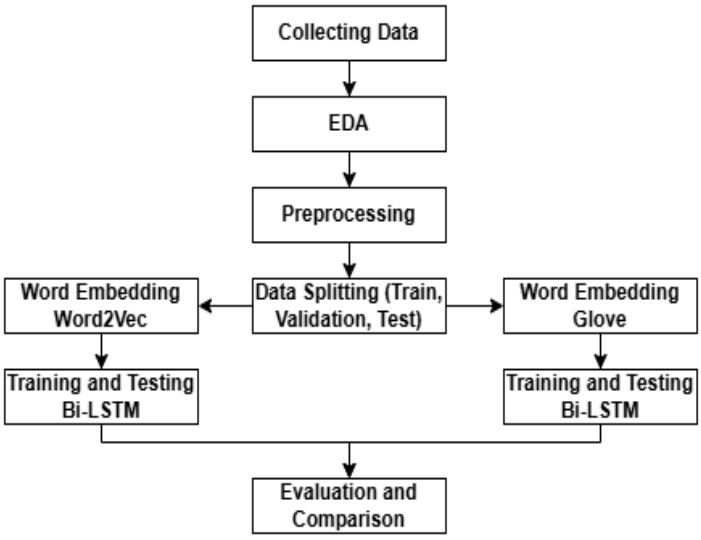

Figure 1. System Architecture

2.1 Data Collection

The first step in data collection was to identify specific data requirements in accordance with the research objectives. In this study, the data used was an Indonesian-language news dataset sourced from the online news portal Kompas.com. This dataset consisted of five news categories labeled sport, travel, politics, education, and money. These news articles were collected through a crawling process using BeautifulSoup (bs4) based on URL or category tag main from the source portal, with a total of 10,000 articles, where each category contains 2,000 news articles.

2.2 Exploratory Data Analysis

In the initial data processing stage, Exploratory Data Analysis (EDA) is an important first step to understand the characteristics of the available data and gain an initial understanding that may influence the selection of further analysis methods [13]. At this stage, the news dataset will be analyzed to identify possible patterns, detect problems or irregularities in the data, and understand the distribution and relationships between variables.

2.2.1 Descriptive Statistics

The collected news dataset consists of four features, namely title, date, content, and category. Statistics on the word length distribution of each news article are attached in Figure 2.

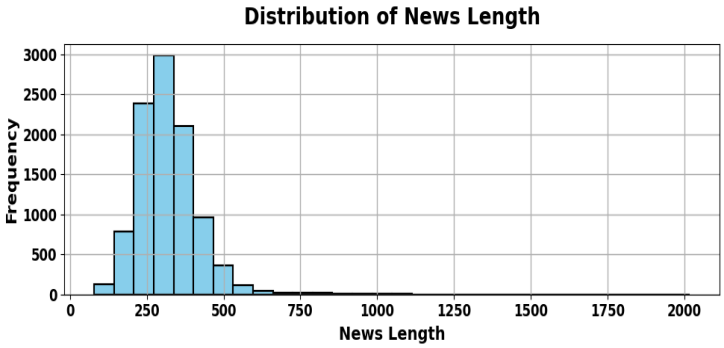

Figure 2. Distribution of Length of Each News Article

Figure 2 illustrates the length distribution of each news article, which is based on the combined 'title' and 'content' features. The graph reveals a significant variation in word count per article, with lengths ranging from 0 to 2000 words. Table 1 presents the dataset that was used.

Table 1. Dataset

Content	Category
investor pasar modal indonesia tembus juta mayoritas di bawah usia tahun pt bursa efek indonesia bei melaporkan bahwa...	Education
bali pas pencinta liburan slow travel liburan merupakan waktu tepat melepaskan diri rutinitas sehari hari...	Travel
cara tarik tunai bri indomaret alfamart mudah nasabah bank rakyat indonesia bri kesulitan mencari mesin atm…	Money
rendahnya partisipasi pemilih pilkada jadi hukuman partai politik data sistem informasi rekapitulasi sirekap komisi pemilihan umum...	Politics
simulasi kredit toyota fortuner gr sport x per bulan rp jutaan toyotafortunergr sport x…	Sport

2.2.2 Data Visualization

To provide a clearer picture of article distribution and word frequency, the data will be visualized using a Word Cloud. A Word Cloud can provide an overview of the frequency of words that frequently appear in each category. The Word Cloud for each category can be seen in Figure 3.

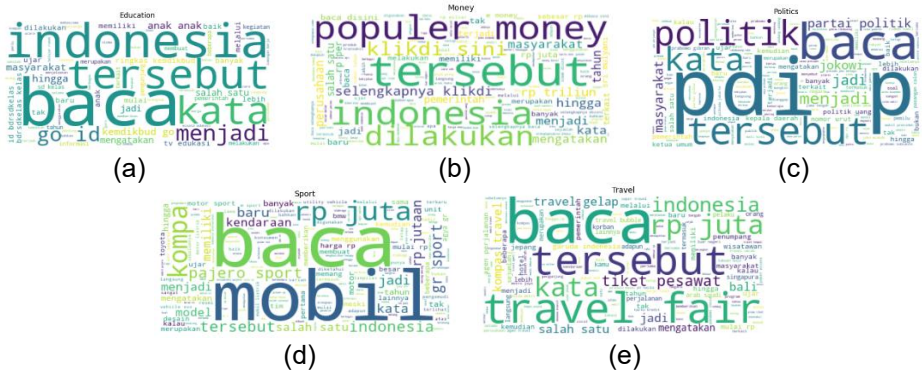

Figure 3. Word Cloud for each Category: (a) Education (b) Money (c) Politics (d) Sports (e) Travel

2.3 Preprocessing

At this stage, text preprocessing is carried out through several crucial steps namely text cleaning, case folding, tokenization, and stopword removal to reduce data noise and normalize the text for more effective feature extraction [14]. The purpose of text cleaning is to remove irrelevant elements from the text that could interfere with the modeling process. At this stage, all punctuation marks and special characters are removed from each news article. Case folding aims to convert all text in the dataset to lowercase letters. This step ensures that the model does not consider words such as "Government", "government", and "GOVERNMENT" as different words. After the text is clean and uniform, the next stage is tokenization. Tokenization breaks each sentence in the news into several words called tokens. Stopwords

Stopwords are common words that appear frequently but have no meaning or information, such as "and," "which," "in," or "is." These words will be removed from the token list because they do not provide significant information.

2.4 Word Embedding

After the data is cleaned and tokenized, the words will be converted into numerical vector representations. This study will implement several modeling scenarios. The embedding models to be tested are Word2Vec with CBOW architecture, Word2Vec with Skip-gram architecture, and GloVe.

2.4.1 Word2Vec

Word2Vec is a word embedding algorithm developed by a Google team led by Tomas Mikolov in 2013 [15]. By using a neural network algorithm that maps words into semantic vectors based on their local context. The CBOW model works by predicting a target word based on the words surrounding it [16]. This model's objective is to predict a target word by utilizing its surrounding context words within a defined window. These context words serve as the input for the prediction. The model's objective is to maximize the probability of the target word's occurrence given its context, as expressed in the following Equation 1:

$$Pr(w_i \mid w_{i-d}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+d}) \quad (1)$$

In this formula, w_i epresents the target word, and d efines the size of its surrounding context window. In practice, this means the model uses the words before and after the target word as input to predict the target word [17].

The Skip-gram model functions as the inverse of the CBOW model. Instead of using the surrounding context to predict a central word, Skip-gram begins with a single target word as its input. The model's primary objective is to then use that input word to predict the words that surround it [18]. In other words, it attempts to forecast the context words that appear in the vicinity of the target word, all within a predefined window.

2.4.2 Global Vector

Global Vector (GloVe) is a word embedding technique introduced by Pennington, Socher, and Manning in 2014 [19]. The key difference from Word2Vec, which only uses local context information, is that GloVe combines local context with global statistics drawn from the enitre dataset. The GloVe algorithm begins by building a global co-occurrence matrix, which stores the frequency of word pairs appearing together within a specific context, aggregated over the entire dataset. This matrix is then used to learn two vectors for each word, one for the context and one for the word itself. During training, the model's aim is to reduce the discrepancy between two values: the dot product of the two vectors and the logarithm of that word pair's probability of occurring together [20].

The difference between GloVe and Word2Vec is in how they learn word relationships. While Word2Vec learns word relationships based on windows, GloVe considers the overall occurrence statistics of words in a document [21]. The fundamental formula for GloVe is shown in Equation 2.

$$w_i^T + \overrightarrow{w_k} + b_i + \overrightarrow{b_k} = \log(X_{ik}) \quad (2)$$

$\overrightarrow{w}$ = Word Context Vector
b_i = bias
b_k = word context bias
X = emergence matrix

The training process for the GloVe model, designed to achieve this goal, involves minimizing the difference between both sides of the above equation for all existing word pairs. This process is formulated into a cost function in the form of weighted least squares regression [22], which is defined in Equation 3.

$$J = \sum_{i,k-1}^{V} f(x_{ik})(w_i^T \overrightarrow{w_k} + b_i + b_k - \log x_{ik})^2 \quad (3)$$

The primary limitation of Word2Vec and GloVe is that they generate static word representations. Consequently, the models cannot handle words that are absent from the training corpus and are unable to distinguish multiple meanings of a single word depending on the context of the sentence [23]. Nevertheless, in computationally constrained scenarios, Word2Vec and GloVe remain crucial to be evaluated as efficient and lightweight baselines for Indonesian NLP tasks.

2.5 Bidirectional Long Short-Term Memory

Traditional LSTM models only process information sequentially in one direction, from the beginning to the end of the sequence. This causes the condition at time t to only depend on information before t and ignore information after it. However, to understand the context of a text as a whole, information from subsequent words is just as important as information from previous words [9]. LSTM stores and accesses information over a long period of time through a cell state mechanism, which is managed by three main gates, namely the Forget Gate, Input Gate, and Output Gate [24].

To capture contextual information more effectively, the Bidirectional Long Short-Term Memory (BiLSTM) model is used. The BiLSTM model is composed of two LSTM networks operating simultaneously, one network processes the sequence from left to right, and the second network processes it from right to left [23]. The architecture of the Bi-LSTM model is depicted in Figure 4.

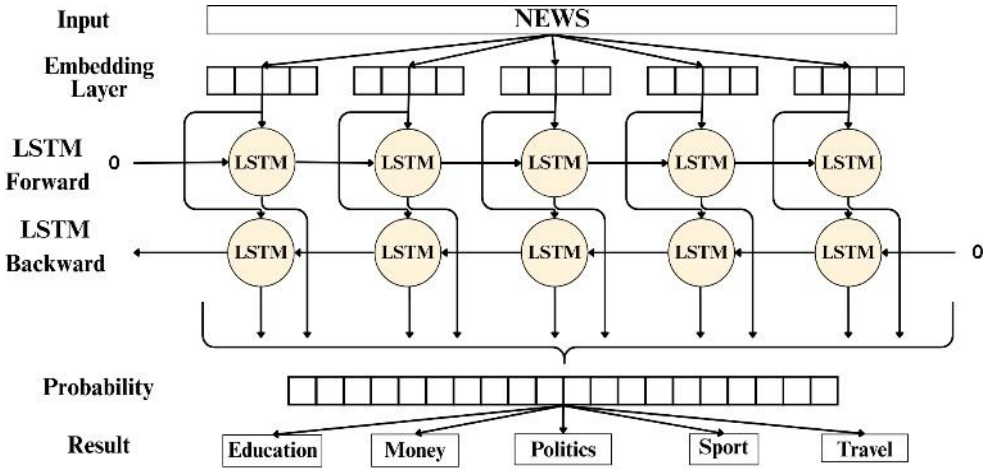

Figure 4. Bi-LSTM Architecture

Consequently, the BiLSTM outputs two hidden states for each time step: one generated from the forward pass and one from the backward pass. The formula for the Forward LSTM pass is shown in Equation 4:

$$\overrightarrow{h_t} = LSTM(x_t, h_{t-1}) \tag{4}$$

The Forward LSTM component processes the sequence from t=1 to t=n to generate a series of forward hidden states, while the backward LSTM processes the sequence in reverse from t=n to t=1 to produce a series of backward hidden states, as formulated in Equation 5.

$$\overleftarrow{h_t} = LSTM\left(x_t, \overleftarrow{h}_{t+1}\right) \tag{5}$$

At time t, the final output of the BiLSTM layer is formed by combining the forward hidden state and the backward hidden state. This ensures that the representation at every time step encapsulates information from the full context. which is mathematically expressed in Equation 6.

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{6}$$

2.6 Evaluation

In this study, the classification models performance is assessed using four key metrics: accuracy, precision, recall, and F1-Score. These are needed to measure each model's effectiveness. By integrating precision and recall, the F1-Score serves as an important metric for classification performance, offering the algorithm an advantage in its overall accuracy and detection abilities. The confusion matrix is an evaluation method that provides a thorough overview of the model's prediction performance. The prediction outcomes in a confusion matrix fall into four main groups, which are: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In this study, the confusion matrix used is a multiclass confusion matrix, which evaluates predictions of more than two classes. Figure 5 shows a multiclass confusion matrix.

		Predicted Class			
		C_1	C_2	...	C_N
Actual Class	C_1	TN	FP	...	TN
	C_2	FN	TP	...	FN
	...	...	...	...	...
	C_N	TN	FP	...	TN

Figure 5. Multiclass Confusion Matrix

Based on the confusion matrix outcomes, the model's performance in predicting the specific type of diagnosis is evaluated using Accuracy, Precision, Recall, and F1-Score, which are mathematically defined in Equations 7, 8, 9, and 10, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{10}$$

2.7 Test Scenario

This study evaluates three modeling scenarios CBOW + Bi-LSTM, Skip-Gram + Bi-LSTM, and GloVe + Bi-LSTM. The sequence length of each article was standardized to 100 tokens, and shorter texts were adjusted using zero padding. Truncating the sequence to 100 tokens was a deliberate experimental design to evaluate the model's robustness and efficiency when provided with limited textual data. This approach tests whether the model can accurately classify articles relying solely on the informative initial segment of the news. In the word representation stage, the Word2Vec models were trained independently using the research dataset with a vector dimension of 100, a context window size of 5, a minimum word frequency of 1, and a training duration of 10 epochs. Meanwhile, the feature representation for the GloVe scenario was obtained by utilizing a 50-dimensional pre-trained Indonesian Wikipedia model. This training process is run for 10 epochs. A summary of the parameters used for the Word2Vec training process can be seen in Table 2.

Table 2. Word2Vec Training Parameters

Parameter	Value	Description
vector_size	100	Dimension of the generated word vector
window	5	Context words surrounding the target word
min_count	1	Ignore all words with a frequency lower than 1 for vocabulary formation
sg	0/1	Training algorithm (0 for CBOW, 1 for Skip-Gram)
epoch	10	Number of training iterations on the dataset

The Bi-LSTM model will use 128 units for the first layer and is set to return the entire sequence so that it can be processed by the next layer. In this study, two dense layers will be implemented, the first using 64 ReLU activation units, while the second will use 5 Softmax activation units to adjust the number of classes in the category . For optimization, the Adam Optimizer will be used with a learning rate of 0.0001 and 10 epochs.

3. Results and Discussion

3.1 Dataset

The dataset consists of news crawled from the kompas.com news portal, totaling 10,000 Indonesian news articles with 2,000 articles per class. After going through the preprocessing stage, the amount of valid data ready for use was reduced to 6,715 news articles. This reduction is the result of two preprocessing steps. First, removing duplicate data to prevent bias in model training. Second, removing outliers based on word length distribution analysis. The analysis results in Figure 2 show that the majority of articles have a length in the range of 100-400 words. Therefore, 2,611 news

articles that fall outside this range were discarded. The distribution of data lengths in the 100-400 word range can be seen in Figure 6.

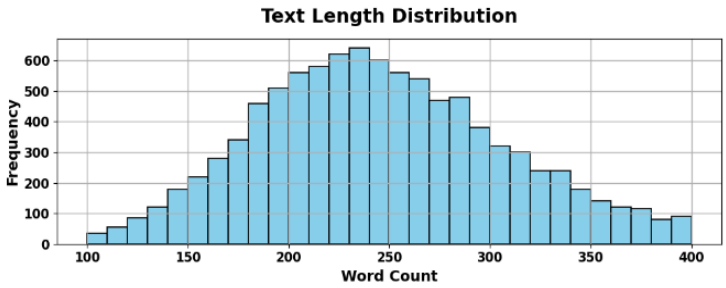

Figure 6. Distribution of Data Lengths

The preprocessing steps resulted in an imbalance in the number of samples across categories. To ensure the model could be trained optimally without bias toward the majority class, a data balancing process was conducted. After evaluation, the Education category became the minority class with the smallest number of samples, namely 1,343 articles. Based on this baseline, the number of data points in the other categories was subsequently reduced to match the Education category. Through this stage, a final dataset with a completely balanced class distribution was obtained, comprising 1,343 articles per category, as shown in Figure 7.

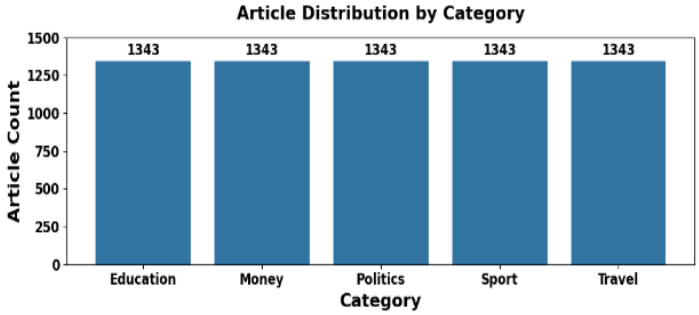

Figure 7. Article Distribution by Category

The cleaned dataset was divided into three distinct sub-datasets: training data, validation data, and testing data. The final performance of the model was evaluated separately using 20% testing data, which was completely unseen during the training phase, to assess the final model's performance on previously unseen data. Meanwhile, the K-Fold Cross-Validation technique was applied to the training data to provide training and validation sets to measure its performance during the development process. The accuracy value reported in the research results represents the model's final performance on the testing data, not the average value from the K-Fold process.

3.2 Experiment Results

The results of the three experimental scenarios, namely CBOW embedding with Bi-LSTM, Skip-Gram embedding with Bi-LSTM, and GloVe embedding with Bi-LSTM, will be presented and analyzed based on accuracy and loss graphs. A confusion matrix will be displayed to see the model's ability to classify each news category.

3.2.1 CBOW + Bi-LSTM Model

The first model tested is a combination of Word2Vec embedding with CBOW architecture and the Bi-LSTM model. The performance of the model during the training and validation processes is visualized through the accuracy and loss curves in Figure 8.

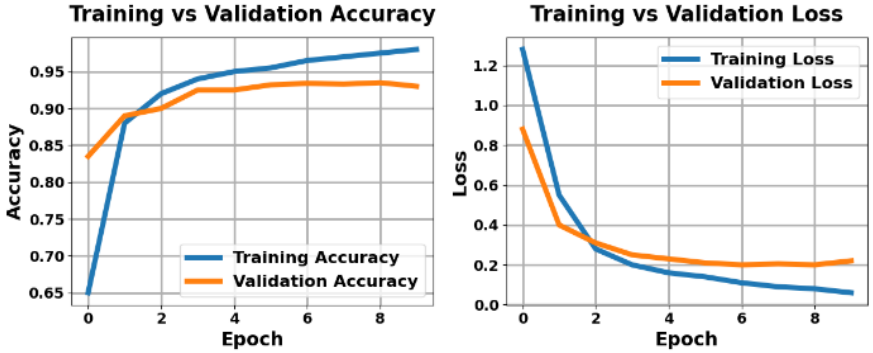

Figure 8. CBOW+BiLSTM Accuracy and Loss Graph

Figure 8 illustrates the CBOW+Bi-LSTM model's performance, showing a consistent training accuracy increase up to 98% and a validation accuracy peaking at 93%, with converging loss values approaching 0.2%. This indicates stable learning without severe overfitting.

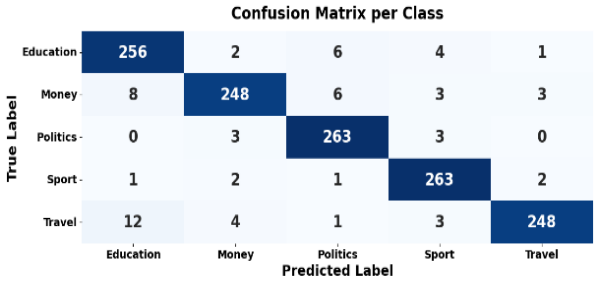

Figure 9. Confusion Matrix CBOW+Bi-LSTM

The confusion matrix in Figure 9 reveals highly precise classification, particularly in Politics and Sports. The model's superior performance stems from CBOW's mechanism of predicting a target word based on its surrounding context. This aligns effectively with the formal and cohesive sentence structures typically found in Indonesian news articles, allowing the model to capture frequent contextual patterns accurately. A minor misclassification occurred where 12 Travel news items were predicted as Education, indicating a degree of vocabulary overlap.

3.2.2 Skip-Gram + Bi-LSTM Model

A combination of the Word2Vec Skip-Gram architecture and the Bi-LSTM model constitutes the second model. Figure 10 provides a visual representation of this model's performance through its accuracy and loss graphs.

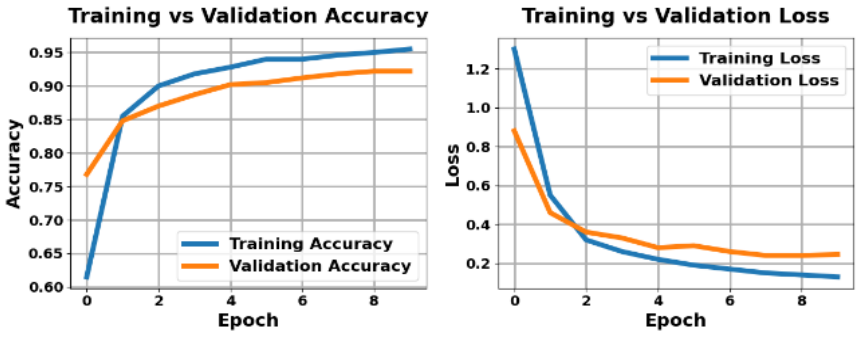

Figure 10. Skip-Gram+BiLSTM Accuracy and Loss Graph

As shown in Figure 10, the Skip-Gram+Bi-LSTM model exhibits a similar but slightly lower performance trajectory, with validation accuracy peaking at 95% and loss stabilizing at 0.25%.

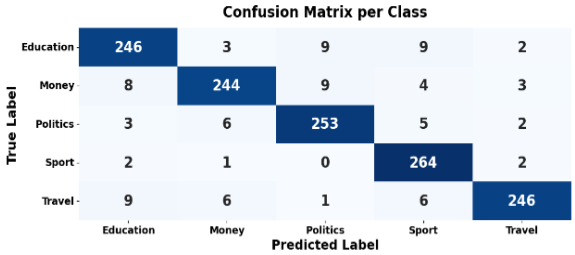

Figure 11. Confusion Matrix for Skip-Gram+Bi-LSTM

While Figure 11 demonstrates strong predictive capabilities, notably in the Sport class (264 correct predictions), the error distribution is more scattered compared to CBOW. For instance, the Education class saw 9 incorrect predictions directed toward both Politics and Sports. Skip-Gram's architecture, which predicts surrounding context from a single target word, excels in representing infrequent words but is slightly less adept than CBOW at synthesizing the comprehensive, smooth contextual flow required for categorizing formal news documents.

3.2.3 GloVe + Bi-LSTM Model

The last model combines pre-trained GloVe embeddings with the Bi-LSTM model. The accuracy and loss graphs for the model can be seen in Figure 12.

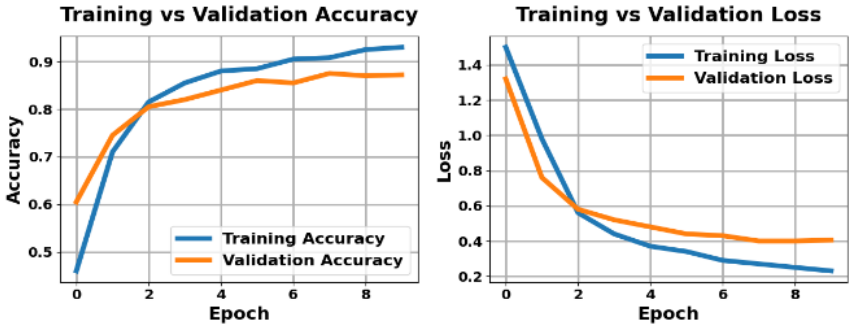

Figure 12. GloVe+BiLSTM Accuracy and Loss Graph

Figure 12 displays a noticeably lower performance for the GloVe+Bi-LSTM model, achieving a peak validation accuracy of 87% and a higher validation loss of 0.4%.

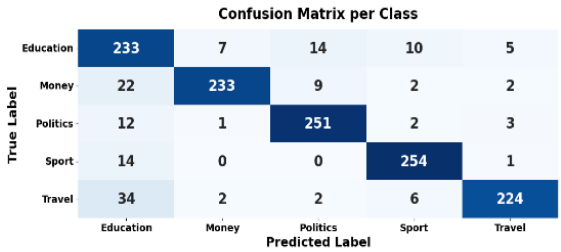

Figure 13. GloVe+Bi-LSTM Confusion Matrix

The confusion matrix in Figure 13 highlights the model's primary weakness: a significant tendency to misclassify various categories into the Education class. This degradation occurs because the pre-trained GloVe embedding relies on global statistical co-occurrences from a general corpus. Consequently, it struggles to distinguish the nuanced, domain-specific boundaries and unique lexical characteristics inherent in this Indonesian news dataset, resulting in less discriminative feature representations compared to the independently trained Word2Vec models.

3.3 Comparison of the Three Models

The comparison of the three embeddings was conducted by observing the accuracy, precision, recall, F1-Score, and computation time during the testing process. The results of the three embeddings can be seen in Table 4.

Table 3. Comparison of CBOW, Skip-Gram, GloVe, and Bi-LSTM Embeddings

Method	Accuracy	Precision	Recall	F1-Score	Time (seconds)
CBOW+Bi-LSTM	95.16%	95.21%	95.16%	95.15%	1.661061s
Skip-Gram+Bi-LSTM	93.30%	93.36%	93.30%	93.29%	1.287479s
GloVe+Bi-LSTM	88.98%	89.75%	88.98%	89.14%	1.614523s

Based on Table 4, the CBOW+Bi-LSTM architecture demonstrates the most superior performance. The key aspect of CBOW+Bi-LSTM is its balance between computational efficiency and precision. Skip-Gram exhibits the fastest computation time, but its accuracy is 2% lower compared to CBOW. This indicates that CBOW's mechanism of combining context to predict a target word is better suited for the formal structure of Indonesian news. Conversely, the GloVe+Bi-LSTM model shows a significant drop in performance. This confirms that pre-trained vector representations based on a general corpus from Indonesian Wikipedia do not effectively capture the unique word choices and vocabulary distribution within the target news domain, especially when compared to a Word2Vec model trained directly on the specific dataset.

3.4 Classification Error Analysis

Although the CBOW+Bi-LSTM model yielded the highest accuracy, the evaluation of the multiclass confusion matrix and per-class evaluation metrics indicated the presence of prediction error patterns in certain classes.

3.4.1 Vocabulary Overlap

Based on the matrix in Figure 13, the model made the most errors by predicting the Travel and Money classes as the Education class. Word frequency distribution analysis revealed a high vocabulary overlap among these classes. For instance, articles in the Travel class frequently contain words such as "educational tourism", "history", and "museum", whereas articles in the Money class often contain terms like "student financial literacy", "scholarship", or "education funds". This feature intersection causes the representation vectors to be close to each other in the embedding space, making it difficult for the model to predict the correct class.

3.4.2 Case Analysis of Classification Errors

Classification model errors occur due to the presence of shared vocabulary across several attributes. For example, in cases where the Travel class is predicted as Education, the text often contains vocabulary from the educational domain, such as 'history', 'education', and 'student'. The Bi-LSTM model fails to capture the sequence properly, resulting in feature weights that are much stronger compared to the actual Travel class. Consequently, the prediction probability shifts and leans more toward the Education class. A similar issue was found in the high rate of prediction errors from the Money class to Education. Much of the news content in the Money class has weights that lean more toward the Education class; for example, in news discussing fiscal policy or government fund allocation, the text may specifically refer to the recipients of the funds, such as schools, scholarships, or students. These tokens can deceive the Word Embedding representation.

3.4.3 The Effect of Article Length on Prediction Errors

The limitation of truncating the text to the first 100 tokens also affects prediction errors. In some articles, the core topic is not placed in the opening paragraph but is obscured by a lengthy introductory narrative, such as history in Travel news or educational background in Money news, causing the model to lose its primary defining context. Despite this limitation, the overall accuracy achieved firmly proves the model's robustness. This indicates that the majority of the first 100 tokens already contain highly representative features, ensuring that the classification remains highly effective even with limited input information.

4. Conclusion

Based on the evaluation results, it can be concluded that the CBOW+Bi-LSTM model demonstrated the best performance, achieving an accuracy of 95.16% and an F1-Score of 95.15%. These results confirm that training embeddings directly on a domain-specific dataset is superior in capturing context for Indonesian news classification. These findings empirically corroborate the research by Lassner, which also proved that generating word embeddings directly from a local corpus consistently outperforms global pre-trained model representations in specialized text classification tasks [25]. The Skip-Gram model followed with solid performance and the fastest computation time, whereas the GloVe+Bi-LSTM model showed the lowest performance due to the incompatibility of general vector representations with news-specific vocabulary.

This study is the initial stage of ongoing research. The primary focus at this stage is to evaluate the robustness of efficient deep learning architectures when operating under strict input data constraints; thus, comparisons with

traditional models such as SVM and Naive Bayes have not been included in the current scope. Although Transformer-based architectures currently offer state-of-the-art performance, these models have drawbacks in the form of massive computational resource requirements and reliance on massive data volumes. In contrast, this study successfully proved that despite being faced with limited word input quantities, models with lighter parameters like Bi-LSTM remain robust and are capable of achieving high levels of accuracy. The results of this study serve as a highly efficient baseline foundation for Indonesian text classification, which will be directly compared with Transformer architectures in future developmental research.

Acknowledgement

The author would like to thank the supervisor of the Informatics Engineering Study Program for all the guidance and input provided.

References

[1] D. Cordeiro, C. Lopezosa, and J. Guallar, "A Methodological Framework for AI-Driven Textual Data Analysis in Digital Media," *Future Internet*, vol. 17, no. 2, Feb. 2025. https://doi.org/10.3390/fi17020059

[2] N. Newman, A. Ross Arguedas, C. T. Robertson, R. Kleis Nielsen, and R. Fletcher, "Reuters Institute Digital News Report 2025". https://doi.org/10.60625/risj-8qqf-jt36

[3] I. Ghozi Zulfikar, Y. Wibisono, and A. Wahyudin, "Intelligent News Aggregation System with Automatic Classification, Clustering, and Summarization," vol. 5, no. 2, 2025. https://doi.org/10.47709/brilliance.v5i2.6712

[4] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," Nov. 01, 2024, *Elsevier Ireland Ltd*. https://doi.org/10.1016/j.cosrev.2024.100664

[5] K. Babić, S. Martinčić-Ipšić, and A. Meštrović, "Survey of neural text representation models," Nov. 01, 2020, *MDPI AG*. https://doi.org/10.3390/info11110511

[6] S. F. Sabbeh and H. A. Fasihuddin, "A Comparative Analysis of Word Embedding and Deep Learning for Arabic Sentiment Classification," *Electronics (Switzerland)*, vol. 12, no. 6, Mar. 2023. https://doi.org/10.3390/electronics12061425

[7] D. S. , N. N. K. & S. P. Asudani, "Impact of word embedding models on text analytics in deep learning environment: a review.," *Artif. Intell. Rev.*, Sep. 2023. https://doi.org/10.1007/s10462-023-10419-1

[8] A. Vallebueno, C. Handan-Nader, C. D. Manning, and D. E. Ho, "Statistical Uncertainty in Word Embeddings: GloVe-V," Jun. 2024. https://doi.org/10.48550/arXiv.2406.12165

[9] H. Alkaabi, A. K. Jasim, and A. Darroudi, "From Static to Contextual: A Survey of Embedding Advances in NLP," *PERFECT: Journal of Smart Algorithms*, vol. 2, no. 2, pp. 57–66, Jul. 2025. https://doi.org/10.62671/perfect.v2i2.77

[10] E. Prasetio Widhi and D. Hatta Fudholi, "Implementation of Deep Learning for Fake News Classification in Bahasa Indonesia," vol. 03, no. 02, pp. 370–381. https://doi.org/10.59141/jrssem.v3i2.546

[11] Z. Li, A. Basit, A. Daraz, and A. Jan, "Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network," *PLoS One*, vol. 19, no. 1 January, Jan. 2024. https://doi.org/10.1371/journal.pone.0291240

[12] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access*, vol. 8, pp. 73992–74001, 2020. https://doi.org/10.1109/ACCESS.2020.2988550

[13] J. Addy, I. Anafo, and E. Westman, "The Role of EDA in Developing Robust Machine Learning Models for Lithology and Penetration Rate Prediction from MWD Data," *Mining*, vol. 6, no. 1, Mar. 2026. https://doi.org/10.3390/mining6010019

[14] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, Mar. 2024. https://doi.org/10.1016/j.is.2023.102342

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013. https://doi.org/10.48550/arXiv.1301.3781

[16] M. Gedeon, "A Comparative Analysis of Static Word Embeddings for Hungarian," May 2025. https://doi.org/10.48550/arXiv.2505.07809

[17] H. A. Almuzaini and A. M. Azmi, "Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020. https://doi.org/10.1109/ACCESS.2020.3009217

[18] G. Sgroi *et al.*, "Evaluation of word embedding models to extract and predict surgical data in breast cancer," *BMC Bioinformatics*, vol. 22, Nov. 2021. https://doi.org/10.1186/s12859-022-05038-6

[19] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation.".

[20] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," Dec. 01, 2019, *Academic Press Inc*. https://doi.org/10.1016/j.yjbinx.2019.100057

[21] A. Nurdin, B. Anggo, S. Aji, A. Bustamin, and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2Vec, Glove, dan Fasttext Pada Klasifikasi Teks," *Jurnal TEKNOKOMPAK*, vol. 14, no. 2, p. 74, 2020. https://doi.org/10.33365/jtk.v14i2.732

[22] M. G. Adrian, S. S. Prasetyowati, and Y. Sibaroni, "Effectiveness of Word Embedding GloVe and Word2Vec within News Detection of Indonesian uUsing LSTM," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 3, p. 1180, Jul. 2023. https://doi.org/10.30865/mib.v7i3.6411

[23] G. F. Shidik *et al.*, "Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM)," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 3, Sep. 2024. https://doi.org/10.1016/j.joitmc.2024.100358

[24] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019. https://doi.org/10.1109/ACCESS.2019.2909919

[25] D. Lassner, S. Brandl, A. Baillot, and S. Nakajima, "Domain-Specific Word Embeddings with Structure Prediction," *Trans. Assoc. Comput. Linguist.*, 2023.

Appendix 1. Per Class Matrix Report

Method	Class	Accuracy	Precision	Recall	F1-Score
CBOW+Bi-LSTM	Education	95%	92%	95%	94%
	Money	93%	96%	93%	94%
	Politics	98%	95%	98%	96%
	Sport	98%	95%	98%	97%
	Travel	93%	98%	93%	95%
Skip-Gram+Bi-LSTM	Education	91%	92%	91%	92%
	Money	91%	94%	91%	92%
	Politics	94%	93%	94%	94%
	Sport	98%	92%	98%	95%
	Travel	92%	96%	92%	94%
GloVe+Bi-LSTM	Education	87%	74%	87%	80%
	Money	87%	96%	87%	91%
	Politics	93%	91%	93%	92%
	Sport	94%	93%	94%	94%
	Travel	84%	95%	84%	89%