

d_Bidirectional_GRU_and_LSTM_Model_on_Indonesian_News_Portal.pdf

by Ahmad Faza

Submission date: 06-Oct-2026 01:57PM (UTC+0700)

Submission ID: 3052015528

File name: d_Bidirectional_GRU_and_LSTM_Model_on_Indonesian_News_Portal.pdf (802.53K)

Word count: 5146

Character count: 28578

Detecting Evergreen Articles: Comparing Word Embeddings for Hybrid Bidirectional GRU and LSTM Model on Indonesian News Portal

³ 1st Adhelio Reyhandro
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
adhelio.reyhandro@student.umn.ac.id

⁷ 2nd Ahmad Faza
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
ahmad.faza@umn.ac.id

Abstract—Automatically detecting evergreen articles in Indonesian news media is a challenge and vital task for sustaining long-term SEO and audience engagement in media industry. This research propose a hybrid parallel deep learning model that hybrids bidirectional GRU and LSTM layers with FastText embeddings, and compare its performance against other word embedding methods, CNN-LSTM and IndoBERT. The experiments utilize a balanced dataset of 5,810 manually labeled article titles from Kompas.com, which underwent preprocessing via case folding, punctuation and number removal, and stemming. Three word embedding techniques, TF-IDF, Word2Vec, and FastText, are employed, with each model trained under consistent settings. The results shows TF-IDF and Word2Vec yield acceptable performance, FastText enhances model robustness. The hybrid BiGRU-BiLSTM model attains a test accuracy of 92.51% with low loss, alth³⁸ IndoBERT attains superior performance with 95.09% accuracy. These findings highlight the effectiveness of word embeddings and hybrid architectures for evergreen article detection. Future work could incorporate attention mechanisms and ensemble methods to enhance classification, offering a promising direction for automated news analysis and supporting strategic decision-making in media organizations.

Keywords—CNN, Evergreen Articles, IndoBERT, RNN, Word Embeddings

I. INTRODUCTION

²⁰ Text classification is an essential task in Natural Language Processing (NLP) with a crucial role in structuring and extracting valuable insights from extensive amounts of textual data [1]. Classical machine learning techniques on text classification—such as news categorization—has been outperformed by deep learning models, from feature learning, scalability, to robustness on dealing noises in data [2]. As one of the sub-categories in news text classification, evergreen articles are timeless content that are not bound by time or specific events and usually serve as educational content for readers, making this news category maintain long-term popularity and continue to attract readers over an extended period [3]. Evergreen articles share its importance and benefits for Search Engine Optimization (SEO) in news media industry. The different characteristics of evergreen articles—compared to timely or trending news articles—led to be read across years and gave long-term traffic instead of only being viewed for days or weeks before being irrelevant to the readers [4]. This small fraction of news category may be written purposely or unintentionally, making evergreen articles unlabeled for further analysis on production, traffic, and search engine index position of this article category.

Deep learning method for news text classification can be applied to replace the manual process of detecting evergreen articles, which was traditionally completed by reviewing and

checking each article titles and labeling them manually in batches at specific published date timeframe that follows the SEO analysis requirements. Although there aren't many research specifically on timeless or evergreen content detection, there are similar approaches within the scope of classifying news articles such as fake news detection and news category classification. Neural network models—for instance Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN)—has been proven to be a great match for multiple tasks within news text classification, where CNNs excel at retrieving local and position-invariant characteristics, while RNN models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are effective at attaining long-range dependencies as well context information [5]. Word embeddings or feature extraction in NLP is also an invaluable part for text classification. Methods such as TF-IDF, Word2Vec and FastText are commonly used from sentiment analysis to news categorization as word embedders for classical machine learning or deep learning models that provides high classification accuracy [6, 7, 8]. Other than CNNs and RNNs that uses word embedding methods for its embedding layer, there are transformer models such as BERT that can understand context of words bidirectionally. IndoBERT—a variation of BERT—has been particularly pre-trained on a large Indonesian corpus, which is fitting for Indonesian language cases from text mining to news text classification [9, 10].

While significant advancements have been made in the classification of news articles, detecting evergreen articles presents unique challenges that existing methods do not fully address [4]. The importance of accurately identifying evergreen content extends beyond SEO advantages, as it plays a crucial role in long-term content strategy and maintaining consistent audience engagement. However, the manual method of identifying these articles is laborious, time-consuming, and subject to human error, highlighting the need for automated solutions. To fill the research gaps, this research aims to evaluate and compare combination of word embeddings and neural network models and IndoBERT, as well proposing a parallel hybrid model of BiGRU and BiLSTM with FastText embeddings on classifying evergreen articles based on their titles. Word embedding methods that will be studied are TF-IDF, Word2Vec, and FastText, whereas the deep learning models being utilized for comparison are CNN-LSTM and BiGRU-BiLSTM. The comparison result of this research intends to offer practical insights of word embedding, neural network, and IndoBERT methods in news text classification, as well as promoting the sub-category of news text classification—that is evergreen articles detection—for a more sustainable and competitive strategy on evergreen article production as well as SEO traffic and content analysis within the media industry.

II. METHODOLOGY

A. Relevant Research

Recent research by [11] investigated classifying timely versus timeless (evergreen) content in the Thai language using deep learning. In this study, CNN-based models integrated with FastText embeddings were applied to a dataset of 600 Facebook articles, with the CNN model augmented by FastText Skip-gram achieving around 90% test accuracy, while Transformer-based approaches such as WangchanBERTa reached approximately 93% accuracy. Complementing these findings, [12] analyzed various hybrid RNN architectures on the BBC News Text Categorization dataset—experimenting with combinations of LSTM and 42 U layers—and found that an RNN+GRU model attained the highest performance (77% accuracy with an F1 score of 0.77), also noting that GloVe embeddings were less effective at capturing subword nuances compared to other methods.

Other studies have advanced hybrid deep learning techniques in related domains. For instance, [13] proposed a hybrid CNN-LSTM model that leveraged supervised FastText embeddings for fake news detection, evaluating its performance on three large-scale datasets totaling over 100,000 articles, and achieving near-perfect accuracy and F1 scores (around 99%). Similarly, [14] employed a weakly supervised framework comparing BiGRU and BiLSTM with TF-IDF embedding for detecting fake news in cyber propaganda from a Twitter dataset of 195,943 tweets (with 1,200 manually labeled), where the BiLSTM variant reached nearly 90% accuracy. Additionally, [15] demonstrated the efficacy of pre-trained transformers by fine-tuning IndoBERT on 9,310 Indonesian healthcare reviews—achieving 96% accuracy and an F1 score of 95%—thus underscoring the potential of deep contextualized models in processing Indonesian text.

Building on these insights, this research proposes a parallel hybrid BiGRU–BiLSTM model that integrates FastText embeddings for detecting evergreen articles on an Indonesian news portal. The study will not only leverage the rich subword-level information provided by FastText but will 9 so compare its performance against other word embedding methods such as TF-IDF and Word2Vec. Furthermore, this approach will benchmark the proposed hybrid model against established architectures like CNN-LSTM as well as IndoBERT—where transformer model such as IndoBERT have the ability to understand context from terms or words within the corpus bidirectionally through attention mechanism [16] as well as widely studied and used for cases on Indonesian texts including problems in the scope of Indonesian news classification [9]—by conducting comprehensive experiments on a 9 rated dataset of Indonesian news article titles, aiming to identify the most effective combination between deep learning architecture and word embedding method for robust evergreen article detection.

B. Article Title Dataset

TABLE I. DATA SAMPLE OF KOMPAS.COM ARTICLE TITLES

Title	Label
Gaji UMR Banjar 2025, Masih Terendah di Jawa Barat (Banjar 2025 Minimum Wage, Still the Lowest in West Java)	0

Title	Label
Apa yang Terjadi Jika Bayi Kekurangan Vitamin D? Berikut 4 Daftarnyaâ€¦ (What Happens If Baby Lacks Vitamin D? Here Are 4 Lists)	1
Pantauan Udara Kemenhub Bantu Atur Lalin Selama Nataru (Ministry of Transportation's Air Monitoring Helps Manage Traffic During Christmas and New Year)	0

The dataset that is used for this research is extracted from the Kompas.com article database from 2023 to 2024. Each of the article titles is labeled as evergreen or non-evergreen manually, resulting in 5,810 balanced rows of data. The dataset consists of title and label columns with samples shown in Table I with translation inside of the parentheses.

C. TF-IDF

TF-IDF—or Term Frequency - Inverse Document Frequency—is a statistical technique used in NLP an 22 information retrieval to assess the significance of terms in a document relative to a collection of documents [17]. TF-IDF will be used to repre 30 text data as embedding matrix, with each row representing a document and each column relates to a word in the vocabulary.

$$TF = \frac{\text{Number of times the term appears in document}}{\text{Total number of terms in document}} \quad (1)$$

$$IDF = \log \left(\frac{\text{Documents in corpus}}{\text{Documents in corpus with term}+1} \right) \quad (2)$$

TF determines how frequently a term occurs in a document, 24 e IDF measures the rarity of a term throughout the corpus. The 17-IDF score is assessed by multiplying TF and IDF, giving importance to terms that appear frequently in a given document but infrequently in the corpus [17].

D. Word2Vec

Word2Vec is a simple two-layer 4 neural network for word embedding and is trained using two approaches, Continuous Bag-of-Words (CBOW), which predicts a word based on its context, and Skip-Gram, which predicts the context of certain a word, as well incorporating subsampling, which reduces the frequency of common words, and negative sampling, which limits computational cost by updating only a subset of the network's weights during training [18]. Word2Vec captures semantic relationships between words, it improved initial weight values and the combination with LSTM algorithm enhanced classification accuracy [19]. In this research, Word2Vec will be used with all deep learning models and be compared for its performance.

E. FastText

FastText—developed by Meta's AI Research—is an efficient library for word representation and sentence classification, built on Word2Vec. It utilizes techniques like Hierarchical Softmax and Negative Sampling for optimized training, and breaks words into character n-grams for better understanding. Supporting 157 languages including Indonesian, 16 Text provides pre-trained word vectors trained using CBOW with position-weights, 300 dimensions, 5-character n-grams, window size of 5, and 10 negatives. This library effectively handles rich morphological languages, accurately representing words even if they're not in the training data [20].

F. CNN-LSTM

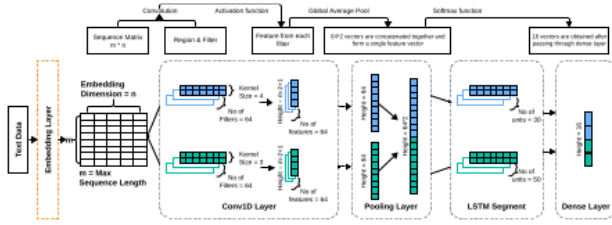


Fig. 1. CNN-LSTM Model Architecture Diagram [13]

The CNN-LSTM model that will be tested in this experiment is a modified version of CNN-LSTM model from [13]. The hybrid model combining CNN and LSTM that is equipped with embeddings from FastText have outperformed other methods in text classification and proven its performance on multiple datasets, with 0.99 classification accuracy on the WELFake and FakeNewsNet dataset. Using the architecture in Figure 1, the model architecture's output layer will be changed from 16 unit softmax activation to 1 unit sigmoid activation, but keeping the L2 kernel regularization used. Other methods to tackle the model from overfitting such as batch normalization and dropouts are also implemented. These architecture modifications are made to ensure the model's compatibility for binary text classification and as a comparison for BiGRU-BiLSTM model.

G. IndoBERT

IndoBERT is configured with the default settings for BERT-Base (uncased), specifically designed for the analysis of Indonesian text. This transformer-based model leverages the BERT architecture by pre-training deep bidirectional representations from unlabeled text across all layers. Trained as a masked language model using the Huggingface framework, IndoBERT effectively utilizes the strengths of BERT. Additionally, this model is capable of performing sequence classification to categorize text based on predefined labels [21]. In this research, IndoBERT will be used as the data tokenizer and be trained on the dataset with classification head by Auto Model For Sequence Classification from Transformers library as performance benchmark for other models.

H. Proposed Model

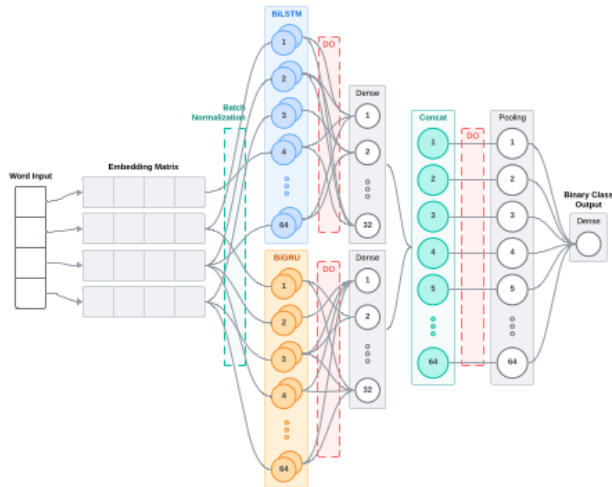


Fig. 2. Proposed Hybrid BiGRU-BiLSTM Model Architecture

Figure 2 presents the neural network diagram of the parallel hybrid BiGRU-BiLSTM model proposed in this experiment, inspired by [22]. It begins with an input layer that processes word sequence based on the max length from the dataset and employs an embedding matrix using FastText, with batch normalization after. LSTM and GRU—each with 64 bidirectional units—receive the input in parallel before entering each respective 32-unit dense layer with ReLU activation and 0.5 dropout to prevent overfitting. These dense layers are concatenated into 64 features and enter the Global Average Pooling layer—which reshapes the 3-dimensional tensor into 2 dimensions—through another 0.5 dropout layer. The model ends with a dense output layer that uses sigmoid activation and L1 regulation of 0.1 for classification results, with the Adam optimizer and binary cross-entropy loss.

I. Research Process Pipeline

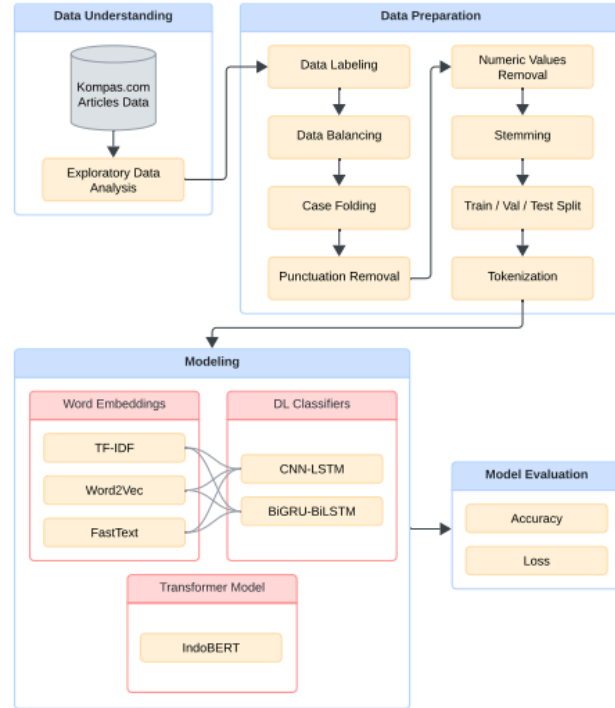


Fig. 3. Research process based of CRISP-DM for comparing text classification methods

Figure 3 presents the process of this research on comparing word embedding methods as well as neural network and IndoBERT models. The process is based of CRISP-DM with adjustments for the needs of this research. Starting from data understanding, the dataset used is explored and prepared for labeling and balancing. Case folding, punctuation and numeric values removal, as well as stemming are also implemented before splitting the data to 80:20 train and test sets, then 80:20 for validation from the train set. Stopwords removal is not used for the case of evergreen detection in this research, as evergreen titles relies heavily in Stopwords such as “cara” (how to), “baik” (good), and “buat” (make). Tokenization used for deep learning models is from Keras, while IndoBERT uses its own model as the tokenizer. Embedding matrixes will be created using TF-IDF, Word2Vec, and FastText to train and pair the embeddings with CNN-LSTM and BiGRU-BiLSTM. Both deep learning models will be trained with 50 epochs and 0.0001 learning rate, while IndoBERT will only be trained with 10 epochs.

After modeling, each model's train validation loss and accuracy will be visualized in addition to printing out the test set loss and accuracy for comparison.

J. Accuracy ¹³

Accuracy is a pivotal metric for evaluating binary classification models, representing the ratio of correct predictions to the total number of predictions. It is valuable and does not mislead if the training uses dataset with balanced label [23].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The formula from [23] encompasses True Positives (TP) and True Negatives (TN)—indicating correct predictions for both positive and negative classes—divided by the sum of all possible outcomes, including False Positives (FP) and False Negatives (FN).

K. Binary Cross-Entropy Loss

Binary Cross-Entropy (BCE) is an important loss function for training binary classification neural networks. It quantifies the discrepancy between actual labels and predicted probabilities by using logarithms of the predictions. Specifically, BCE calculates the loss for each sample and then averages it over all samples to determine the overall loss [24].

$$L = -\frac{1}{n} \sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \quad (4)$$

The simplified formula from [24] includes terms $\log(p_i)$ and $\log(1-p_i)$, which dominate respectively when the true label y_i is 1 or 0, ensuring higher predicted probabilities for the true class result in lower loss values. The negative sign ensures the overall loss is positive. This loss function guides the neural network in adjusting weights during training to enhance prediction accuracy.

III. EXPERIMENT AND RESULT

A. Experiment

The experiment starts by exploring the dataset extracted from the Kompas.com article database. The data have max title length of 29 words with mean of 10 words. Each row of title labeled manually for indications of evergreen content and balanced between the class, resulting in 2,905 of evergreen titles and 2,905 of non-evergreen titles. Case folding is applied, as well removing punctuations and numbers with regular expression. For stemming, Sastrawi library is used as it provides stemmer for Indonesian language. Sample of before and after preprocess for the data is shown in Table II.

TABLE II. DATA SAMPLE OF BEFORE AND AFTER PREPROCESS

Before Preprocess	After Preprocess
Gaji UMR Banjar 2025, Masih Terendah di Jawa Barat	gaji umr banjar masih rendah di jawa barat
Apa yang Terjadi Jika Bayi Kekurangan Vitamin D? Berikut 4 Daftarnya	apa yang jadi jika bayi kurang vitamin d ikut daftar
Pantauan Udara Kemenhub Bantu Atur Lalin Selama Nataru	pantau udara kemenhub bantu atur lalin selama nataru

After preprocess, the dataset is split to training and testing set, then the training set is split to training and validation set, resulting in 3,718 train samples, 930 validation samples, and

1,162 test samples. Splitting for the validation set ensure that each model training have the same train and validation set. Tokenization is performed using Keras tokenizer for the deep learning models, while IndoBERT uses AutoTokenizer with its own model as the tokenizer. The data is ready to be used as embedding matrixes for each word embedding methods. For TF-IDF and Word2Vec, this experiment will use 10,000 max features or words, 100 embedding dimensions, and allows the embedding layer to be trainable, while FastText will use its default 300 dimensions with non-trainable embedding layer to keep the pre-trained weights. Each model will be trained for 50 epochs with $1e-3$ learning rate and batch size of 16, while IndoBERT will be trained for 10 epochs with $1e-5$ learning rate as it uses auto model for sequence classification for transformer models.

B. Result

Each training loss and accuracy results of word deep learning and IndoBERT models, results are visualized.

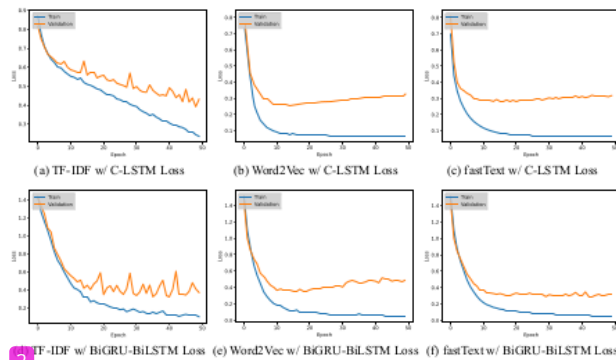


Fig. 4. Training and Validation Loss of the deep learning models

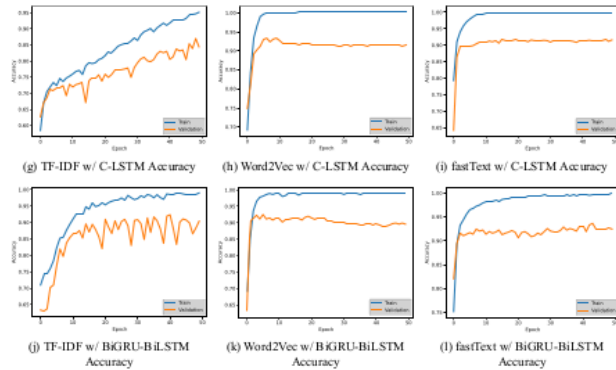


Fig. 5. Training and Validation Accuracy of the deep learning models

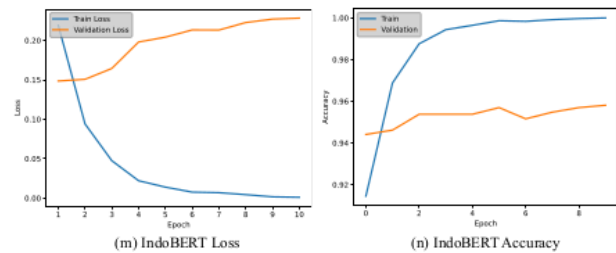


Fig. 6. Training Validation Loss and Accuracy of IndoBERT model

Figure 4 illustrates the training and validation loss results from each pairing between word embedding methods and

deep learning models, where both models with TF-IDF shows the most steady decrease of loss on training and validation set for each epoch with fluctuations. Word2Vec and FastText have stable training and validation loss in both models, however the training loss decreases rapidly and stabilizes near zero while the validation loss stabilizes and start to increase or overfitting after around 15th epoch. FastText with BiGRU-BiLSTM shows no increase of validation loss and plateaus around 0.3000. In Figure 5, TF-IDF on both models trains with fluctuations and steady increase. Both models validation accuracy with Word2Vec and FastText peaks around 92% to 93% while the training accuracy stabilizes near 100%. IndoBERT starts with low loss and high accuracy as shown in Figure 6. However, IndoBERT's validation loss keeps increasing in each epoch, while the validation accuracy plateaus around 95% as the train accuracy already stabilizing near its peak.

TABLE III. MODELS' TEST DATA LOSS

Classification Models / Word Embeddings				
Loss	TF-IDF	Word2Vec	FastText	IndoBERT
CNN-LSTM	0.4006	0.3128	0.3039	-
BiGRU-BiLSTM	0.3349	0.4326	0.3137	-
IndoBERT	-	-	-	0.2603

TABLE IV. MODELS' TEST DATA ACCURACY

Classification Models / Word Embeddings				
Accuracy	TF-IDF	Word2Vec	FastText	IndoBERT
CNN-LSTM	86.14%	92.51%	91.14%	-
BiGRU-BiLSTM	91.99%	90.79%	92.51%	-
IndoBERT	-	-	-	95.09%

Among the various models from the test loss and accuracy pivot tables shown in Table III and Table IV, the FastText with BiGRU-BiLSTM combination [33] nonstrated significant efficacy, highlighting a low test data loss of 0.3137 and a high accuracy rate of 92.51%, while the pairing with TF-IDF gives an adequate results of 0.3346 test data loss as well 91.99% accuracy, and least performed with Word2Vec, 0.4326 test data loss and 90.79% accuracy. Alternatively, the CNN-LSTM models unveiled comparable performance, with the FastText-based model achieving a test data loss of 0.3039 and an accuracy of 91.14%, while the pairing with Word2Vec showed a slightly better accuracy of 92.51% but a higher loss at 0.3128, and performed worse with TF-IDF, showing the lowest accuracy of 86.14% and second to high [6] test loss of 0.4006. Lastly, the IndoBERT model stood out with the lowest test data loss of 0.2603 and the highest accuracy of 95.09%, highlighting its superior capability in capturing the nuanced features of evergreen news articles.

TABLE V. THREE BEST MODELS TEST SET PREDICTION SAMPLE

Title	True Label	Word2 Vec w/ CNN-LSTM	Fast Text w/ BiGRU-BiLSTM	Indo BERT
Film Home Sweet Loan Pamit dari Bioskop,	0	0	1	0

Title	True Label	Word2 Vec w/ CNN-LSTM	Fast Text w/ BiGRU-BiLSTM	Indo BERT
Kantongi Penonton (Home Sweet Loan Movie Leaves Cinemas, Bags 1,720,271 Viewers)				
Ini 4 Rute Angkot Mikrotrans yang Beroperasi 24 Jam (These are 4 Mikrotrans City Transport Routes that Operate 24 Hours)	0	1	0	0
Sederet Efek Samping Obat Penurun Darah Tinggi seperti Amlodipine (A Series of Side Effects of High Blood Pressure Medications such as Amlodipine)	1	1	1	1
Apakah Puasa Bisa Menurunkan Berat Badan? (Can Fasting Help You Lose Weight?)	0	0	1	1
30 Ucapan Jumat Agung dalam Bahasa Indonesia dan Bahasa Inggris (30 Good Friday Greetings in Indonesian and English)	0	1	0	1

Each best model illustrates distinct patterns of errors, reinforcing that different architectures and embedding methods capture textual features in varying ways by looking from the test set prediction samples in Table V. Even though IndoBERT generally achieves higher accuracy overall, it still can mislabel certain titles that contains seasonal events such as "Jumat Agung" or Good Friday, and "Puasa", or fasting season in Ramadhan. Non-transformer models—CNN-LSTM and BiGRU-BiLSTM—also exhibit strengths in some cases but can underperform on more nuanced headlines.

IndoBERT outperforms other models by combining an Indonesian-specific pre-trained tokenizer with the transformer's attention mechanism, enabling it to weigh words contextually and identify evergreen articles [40] precisely. CNN-LSTM with Word2Vec captures local features through convolutional layers and long-term dependencies via LSTM units, and while integrating FastText embeddings improves its results, it still lags behind bidirectional and attention-based architectures. The TF-IDF with CNN-LSTM approach fares worst, underscoring TF-IDF's inability to capture deeper textual nuances crucial for this classification task. Our proposed BiGRU-BiLSTM with FastText model also demonstrates strong performance as its bidirectional GRU and LSTM layers effectively learn sequential patterns from both directions. However, lacking an attention mechanism like IndoBERT's, it peaks at around 0.300 loss and 92% accuracy during training, explaining the performance gap between the hybrid deep learning models with a transformer model such as IndoBERT.

IV. CONCLUSION

The experimental results underline that while embedding techniques such as TF-IDF and Word2Vec yield acceptable performance in detecting evergreen articles, the FastText-based models consistently outperform both embedding methods, with the proposed BiGRU-BiLSTM achieving a robust test loss of 0.3137 and an accuracy of 92.51%. The CNN-LSTM models shown competitive results with FastText and Word2Vec, with its best pair with Word2Vec gives same

accuracy of 92.51% but a slightly lower loss of 6.3128. Notably, IndoBERT emerged as the top performer with the lowest test loss of 0.2603 and the highest accuracy of 95.09%, primarily due to its attention mechanism and language-specific pre-training, which enable it to capture 36 the contextual nuances in Indonesian news headlines. These findings highlight the importance of sophisticated embedding techniques and deep contextual understanding in text classification tasks. Future research should explore integrating attention layers into hybrid RNN frameworks to bridge the performance gap between transformer-based and non-transformer models. Additionally, extending the dataset to incorporate a wider spectrum of news sources and experimenting with ensemble approaches that combine transformer-based insights with traditional deep learning architectures could further enhance classification performance. Refinement of hyperparameters, along with investigating alternative preprocessing techniques and multi-modal features, also presents promising avenues to improve evergreen article detection in diverse real-world applications.

2

ACKNOWLEDGMENT

This research was supported by the Research and Community Service Institution and Student Development at Universitas Multimedia Nusantara. We thank our colleagues at the Big Data Laboratory, Department of Information Systems, for their insights and support, which contributed to the study's success. We appreciate the encouragement and feedbacks from the broader academic community at Universitas Multimedia Nusantara, that have enriched this work. We also sincerely thank OVAL at KG Media in Kompas Gramedia Group, as their support have made this research possible by upbrining this business case and providing the dataset of evergreen article detection.

REFERENCES

- [1] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022. doi:10.1145/3495162
- [2] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2021. doi:10.1145/3439726
- [3] A. Ramadhan, I. Dimiyati, Y. Yusanto, A. Muldi, and R. N. Kurniawati, "Radar Banten's strategy in facing the era of media digitalization," *Jurnal Ilmiah Global Education*, vol. 5, no. 1, pp. 44–53, Mar. 2024. doi:10.55681/jige.v5i1.2366
- [4] Y. Liao, S. Wang, E. H. Han, J. Lee, and D. Lee, "Characterization and Early Detection of Evergreen News Articles," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds., *Lecture Notes in Computer Science*, vol. 11908. Cham: Springer, 2020, pp. 566–582. doi:10.1007/978-3-030-46133-1_33
- [5] I. K. Sastrawan, I. P. A. Bayupati, and D. M. Arsa, "Detection of fake news using deep learning CNN-RNN based methods," *ICT Express*, vol. 8, no. 3, pp. 396–408, Sep. 2022. doi:10.1016/j.icte.2021.10.003
- [6] L. Efizoni, S. Defit, M. Tajuddin, and A. Anggawan, "Komparasi ekstraksi FITUR Dalam Klasifikasi Teks multilabel Menggunakan Algoritma machine learning," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022. doi:10.30812/matrik.v21i3.1851
- [7] M. Umer, Z. Imtiaz, M. Ahmad, M. Nappi, C. Medaglia, G. S. Choi, and A. Mehmood, "Impact of convolutional neural network and FastText embedding on text classification," *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 5569–5585, Aug. 2022. doi:10.1007/s11042-022-13459-x
- [8] Y. Gunawan, J. C. Young, and A. Rusli, "FastText word embedding and random forest classifier for user feedback sentiment classification in Bahasa Indonesia," *Ultimatics: Jurnal Teknik Informatika*, vol. 13, no. 2, pp. 101–107, Jan. 2022. doi:10.31937/ti.v13i2.2124
- [9] S. Isa, G. Nico, and M. Permana, "IndoBERT for Indonesian Fake News Detection," *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, 2022. doi:10.24507/iceel.16.03.289
- [10] B. Richardson and A. Wicaksana, "Comparison of IndoBERT-Lite and RoBERTa in Text Mining for Indonesian Language Question Answering Application," *International Journal of Innovative Computing, Information and Control*, vol. 18, no. 6, pp. 1719–1734, 2022. doi:10.24507/ijicic.18.06.1719
- [11] P. Gatchalee, S. Waijanya, and N. Promrit, "Thai Text Classification Experiment using CNN and Transformer Models for Timely-Timeless Content Marketing," *ICIC Express Letters*, vol. 17, no. 1, pp. 91–101, 2023. doi:10.24507/iceel.17.01.91
- [12] P. Sunagar and A. Kanavalli, "A hybrid RNN based deep learning approach for text classification," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022. doi:10.14569/ijacsa.2022.0130636
- [13] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," in *IEEE Access*, vol. 12, pp. 44462–44480, 2024, doi:10.1109/ACCESS.2024.3381038
- [14] L. Syed, A. Alsaedi, L. A. Alhuri, and H. R. Aljohani, "Hybrid weakly supervised learning with deep learning technique for detection of fake news from Cyber Propaganda," *Array*, vol. 19, p. 100309, Sep. 2023. doi:10.1016/j.array.2023.100309
- [15] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 113–117, 2023. doi:10.14569/IJACSA.2023.0140813
- [16] D. Hu, "An Introductory Survey on Attention Mechanisms in NLP Problems," in Y. Bi, R. Bathia, and S. Kapoor, Eds., *Intelligent Systems and Applications (IntelliSys 2019)*, vol. 1038, pp. 432–448, Aug. 2019, Springer, Cham. doi:10.1007/978-3-030-29513-4_31
- [17] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021. doi:10.11591/eei.v10i5.3157
- [18] G. Di Gennaro, A. Buonanno, A. Di Girolamo, A. Ospedale, F. A. Palmieri, and G. Fedele, "An analysis of word2vec for the Italian language," *Smart Innovation, Systems and Technologies*, pp. 137–146, Jul. 2020. doi:10.1007/978-981-15-5093-5_13
- [19] R. Indra and A. S. Girsang, "Classification of User Comment Using Word2vec and Deep Learning," in *Advances in Science, Technology and Engineering Systems Journal (ASTESJ)*, vol. 6, no. 2, pp. 566–572, Mar. 2021, doi:10.25046/aj060264
- [20] "fastText," [Online]. Available: <https://fasttext.cc/>. [Accessed: 4-Feb-2025]
- [21] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, 2023. doi:10.20473/jisebi.9.2.253-263
- [22] A. S. Girsang and Stanley, "Hybrid LSTM and GRU for Cryptocurrency Price Forecasting Based on Social Network Sentiment Analysis Using FinBERT," in *IEEE Access*, vol. 11, pp. 120530–120540, 2023, doi:10.1109/ACCESS.2023.3324535
- [23] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "The Accuracy Comparison Among Word2Vec, GloVe, and FastText Towards Convolution Neural Network (CNN) Text Classification," in *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 2, pp. 349–359, Jan. 2022. [Online]. Available: <https://www.jatit.org/volumes/Vol100No2/5Vol100No2.pdf>. [Accessed: 20-Feb-2025]
- [24] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," in *IEEE Access*, vol. 8, pp. 4806–4813, 2020, doi:10.1109/ACCESS.2019.2962617

ORIGINALITY REPORT

14%

SIMILARITY INDEX

10%

INTERNET SOURCES

10%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1	Fikri Baharuddin, Mohammad Farid Naufal. "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy", Journal of Information Systems Engineering and Business Intelligence, 2023 Publication	1%
2	Fiorenza Irene Christabelle Suryanto, Samuel Ady Sanjaya. "Forecasting of Clinical Program Gross Growth over Revenue through Outcomes and Workforce Metrics using LSTM-XGBoost", 2025 4th International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS), 2025 Publication	1%
3	Valenzia Tanaka, Samuel Ady Sanjaya. "Integration of K-Means and SVM for Customer Segmentation Based on Customer Behaviour Using LRFM and Product Preference", 2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2024 Publication	1%
4	csimq-journals.rtu.lv Internet Source	1%
5	www.techscience.com Internet Source	1%
6	ejournal.nusamandiri.ac.id Internet Source	<1%
7	Zevanaya Beverly Drew, Johan Setiawan, Erick Fernando, Ahmad Faza, Melissa Indah Fianty, Winanti Winanti, Dina Fitria Murad. "Interactive 360° Virtual Tour for Dark Tourism Promotion: A Digital Heritage Case Study of Makam Syekh Mubarak", 2025 8th International Conference on New Media Studies (CONMEDIA), 2025 Publication	<1%
8	ejurnal.stmik-budidarma.ac.id Internet Source	<1%
9	ijece.iaescore.com Internet Source	<1%
10	Submitted to CSU, San Jose State University Student Paper	<1%

11	Sushil Kamboj, Pardeep Singh Tiwana. "Innovations in Computing", CRC Press, 2025 Publication	<1 %
12	stax.strath.ac.uk Internet Source	<1 %
13	Submitted to University of Colorado, Colorado Springs Student Paper	<1 %
14	ris.cdu.edu.au Internet Source	<1 %
15	Nurulhuda Binti Mohamad Ali, Hui-Ngo Goh, Amy Hui-Lan Lim, Nurafiq Inani Man. "MPOS: Development of Malay Part-of-Speech Tagger using Embedding Methods and Neural Network Models", 2025 IEEE 6th International Conference on Pattern Recognition and Machine Learning (PRML), 2025 Publication	<1 %
16	Vasileios Barzokas, Eirini Papagiannopoulou, Grigorios Tsoumakas. "Studying the Evolution of Greek Words via Word Embeddings", 11th Hellenic Conference on Artificial Intelligence, 2020 Publication	<1 %
17	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
18	dergipark.org.tr Internet Source	<1 %
19	etd.hu.edu.et Internet Source	<1 %
20	fastercapital.com Internet Source	<1 %
21	jurnal.iaii.or.id Internet Source	<1 %
22	beei.org Internet Source	<1 %
23	bora.uib.no Internet Source	<1 %
24	cdn.juris.id Internet Source	<1 %
25	ijritcc.org Internet Source	<1 %

26	journal.sah.co.id Internet Source	<1 %
27	www.ijirset.com Internet Source	<1 %
28	Muhammad Syahriandi Adhantoro, Faris Athoil Haq, Dody Hartanto, Aninditawidagda Pandam Sudaryanto. "IndoBERT-Based Sentiment Analysis of Electric Motorcycle Policy in Indonesia Using Instagram Data", Jurnal Penelitian Sains Teknologi, 2026 Publication	<1 %
29	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Computational Intelligence, Information Technology, and Networking - Volume 2", CRC Press, 2026 Publication	<1 %
30	jtiulm.ti.ft.ulm.ac.id Internet Source	<1 %
31	link.springer.com Internet Source	<1 %
32	ojs.imeti.org Internet Source	<1 %
33	su.diva-portal.org Internet Source	<1 %
34	turcomat.org Internet Source	<1 %
35	www.kompas.com Internet Source	<1 %
36	Aide Aracely Maldonado-Macías, Yordán Rodríguez, Juan Luis Hernández Arellano. "Human Factors and Ergonomics in Latin American Countries - Developments, Applications and Approaches", CRC Press, 2026 Publication	<1 %
37	Guo, Danfeng. "Applying Medical Language Models to Medical Image Analysis", University of California, Los Angeles, 2024 Publication	<1 %
38	Hanum Fadiannur, Yuliant Sibaroni. "Exploration of the Effect of Word Embedding Dimension Size: Word2Vec, FastText, and BERT on the Modified BiLSTM Model for Sentiment Analysis of Public Service Applications", 2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), 2025 Publication	<1 %

39	M. Khairul Anam, Sarjon Defit, Rini Sovia. "Evaluating Word2Vec and FastText Embeddings in CNN-LSTM for Hotel Customer Satisfaction: Impact on Accuracy and Model Stability", 2025 International Conference on Data Science and Its Applications (ICoDSA), 2025 Publication	<1 %
40	Muhammad Fairuzabadi, Kusrini Kusrini, Ema Utami, Arief Setyanto. "Advancements and Challenges in Gated Recurrent Units (GRU) for Text Classification: A Systematic Literature Review", 2024 7th International Conference of Computer and Informatics Engineering (IC2IE), 2024 Publication	<1 %
41	arxiv.org Internet Source	<1 %
42	cronfa.swansea.ac.uk Internet Source	<1 %
43	hdl.handle.net Internet Source	<1 %
44	thesai.org Internet Source	<1 %
45	David Olatunde Olukanni, Oluseyi Olanrewaju Ajayi, Joshua Olusegun Okeniyi, Shalom Nwodo Chinedu et al. "Emerging Trends in Sustainable Energy, Smart Systems, and Green Technology", CRC Press, 2026 Publication	<1 %
46	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniques and Smart Manufacturing", CRC Press, 2026 Publication	<1 %

Exclude quotes

On

Exclude matches

< 7 words

Exclude bibliography

On

FINAL GRADE

GENERAL COMMENTS

/0

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6