IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.4, October 2025, pp. 361~370
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.106608

Classifying Indonesian Hoax News Titles with SVM, XGBoost, and BiLSTM

I Nyoman Prayana Trisna*[1], I Made Wiraharja Jaya Putra[2], Wayan Oger Vihikan[3]
[1, 2, 3]Udayana University, Bali, Indonesia
e-mail: ***[1]prayana.trisna@unud.ac.id**, [2]wiraharjajaya@student.unud.ac.id,
[3]oger_vihikan@unud.ac.id

Abstrak

Penelitian ini meneliti deteksi otomatis berita hoaks terkait Presiden Jokowi dalam berita berbahasa Indonesia dengan menganalisis judul berita saja, yang mana bertujuan untuk deteksi yang efisien dan pengurangan traffic ke situs web penyebar hoaks. Kami membandingkan kinerja algoritma machine learning tradisional (SVM, XGBoost) dan deep learning (BiLSTM), dengan dan tanpa teknik oversampling SMOTE untuk mengatasi ketidakseimbangan kelas dalam kumpulan data yang dikumpulkan dari sumber berita terpercaya (CNN Indonesia, Detik News) dan platform pengecekan fakta (turnbackhoax.id). Hasil penelitian menunjukkan bahwa BiLSTM mengungguli SVM dan XGBoost. Namun, penerapan SMOTE berdampak negatif pada kinerja BiLSTM, yang menunjukkan potensi overfitting pada data oversampling. Kemudian, presisi secara konsisten lebih tinggi daripada recall di semua model, yang mengindikasikan kemampuan baik dalam mengidentifikasi judul berita sebagai hoaks tetapi adanya potensi untuk melewatkan sebagian berita hoaks. Penelitian ini menyoroti trade-off antara menghindari false positive dan memastikan deteksi yang komprehensif. Temuan ini juga menunjukkan bahwa karakteristik khusus dalam bahasa memengaruhi efektivitas algoritma. Penelitian ini berkontribusi pada pengembangan model dan alat yang efisien dan akurat untuk memerangi penyebaran informasi yang salah di Indonesia, menekankan pentingnya analisis berbasis judul dan pertimbangan yang cermat terhadap penyeimbangan data.

***Kata kunci**—Deteksi hoax, SVM, XGBoost, BiLSTM, SMOTE*

Abstract

This study investigates the automated detection of hoaxes related to President Jokowi in Indonesian news by analyzing only news titles, aiming for efficient detection and reduced traffic to harmful websites. We compared the performance of traditional (SVM, XGBoost) and deep learning (BiLSTM) algorithms, with and without Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in a dataset scraped from trusted news sources (CNN Indonesia, Detik News) and a fact-checking platform (turnbackhoax.id). The results indicate that BiLSTM generally outperformed SVM and XGBoost, demonstrating the potential of deep learning for this task. However, applying SMOTE negatively impacted BiLSTM's performance, suggesting overfitting. Notably, precision consistently exceeded recall across all models, indicating high reliability in identifying hoaxes but a potential for missing a significant number of actual hoaxes. This highlights a trade-off between avoiding false positives and ensuring comprehensive detection. The findings also suggest that language-specific characteristics influence algorithm effectiveness. This research contributes to developing efficient and accurate tools for combating misinformation in the Indonesian online environment, emphasizing the importance of title-based analysis and careful consideration on data balancing.

***Keywords—** Hoax detection, SVM, XGBoost, BiLSTM, SMOTE*

Received May 9th,2025; Revised October 9th, 2025; Accepted October 27th, 2025

1. INTRODUCTION

Hoaxes are a significant issue in Indonesia. It is a false information that can be understood in different ways by different people. Surveys have shown that these hoaxes are frequently used as political tools [1]. A notable example is their usage against President Joko Widodo, the president of Indonesia from 2014 to 2024. These hoaxes often zero in on sensitive topics such as ethnicity, religion, race, and the relationships between different groups in society. The clear aim is to damage the government's reputation and make the public loses trust [2].

One of the main problems with hoaxes is that they can be very convincing [3]. Because of this, it's crucial to have strong systems in place that can classify them accurately. Previous research has explored various methods to automate this process. This includes using machine learning techniques like Support Vector Machine (SVM) [4], K-Nearest Neighbour [5], and Extreme Gradient Boosting (XGBoost) [6] to aid for automation in hoax news detection. Similarly, researchers have also looked at deep learning methods for such as Bidirectional Long Short-Term Memory (BiLSTM) [7], Convolutional Neural Network (CNN) [8], ResNet [9], and even using pre-trained transformer models [10] or models that use transfer learning [11] to build models which can differentiate between actual or misleading news.

However, a common approach in existing studies is to analyze the entire content of a news article. This method, while providing a lot of information, has a downside: it can unintentionally drive more traffic to the very websites that are spreading these hoaxes. Our research proposes a different way of tackling this. We suggest focusing solely on the titles of news articles to detect hoaxes. The idea behind this is to make the detection process more efficient, reduce the number of people visiting harmful sites, and better align with how users typically read news online – often by scanning headlines first.

The classification of news as a hoax has been a well-researched area, with many different methods being tried. For instance, Cahyo et al. [12] did a comparison of how well LSTM, SVM, and Multinomial Naïve Bayes could categorize hoax news. They used a dataset from a directory that checks facts and found that SVM was more accurate than LSTM. This suggests that there's still room to improve how deep learning models are used for this specific task. Following this, other studies have looked at using SVM to identify hoaxes related to Covid-19. Research by Putri et al. [13] and Ropikoh et al. [14] developed automated systems based on SVM. These systems were tested using k-fold cross-validation and showed good results in terms of precision, recall, and f1-score. However, a limitation in these studies is that they relied on the full text of news articles, which, as mentioned, could lead to more traffic for hoax spreaders. This highlights the need to explore other methods, like classifying based on titles alone, to avoid this issue.

In another line of research, Khanam et al. [15] compared the performance of several machine learning algorithms, including Linear Regression, Random Forest, XGBoost, Naïve Bayes, K-Nearest Neighbors (K-NN), Decision Tree, and SVM. Their findings indicated that XGBoost was quite effective, achieving accuracy rates above 75%. However, their main focus was on traditional machine learning techniques, without deeply exploring the potential of more advanced deep learning models. Their research also suggested that using a combination of models, like XGBoost, might perform better than using single models like SVM. Further studies, such as the one by Haumahu et al. [16], specifically looked at using XGBoost for classifying Indonesian hoax news and found promising results, showing that this algorithm can be effective in this context.

When it comes to deep learning, Aji and Setiawan [17] investigated Recurrent Neural Networks (RNN) and BiLSTM, as well as combining them in hybrid models like RNN-BiLSTM and BiLSTM-RNN. While the performance of the different models varied, the BiLSTM model consistently showed strong performance. In fact, they found that the best results were obtained when using just the BiLSTM model. Similarly, research by Sastrawan et al. [18] demonstrated that BiLSTM is effective in detecting hoax news, even outperforming CNN and

ResNet across multiple datasets. Despite these findings, it's important to note that much of the research has been done on English language datasets. This points to a need for more development of hoax detection systems specifically for Indonesian, taking into account the unique ways hoaxes are written and spread in Indonesia.

To summarize, our current research aims to develop a model that can classify hoaxes related to President Jokowi by only analyzing the titles of news articles. We will compare the performance of SVM, XGBoost, and BiLSTM to see which method works best, as previous studies have shown good results with these models. Ultimately, the goal is to create a tool that can automatically detect hoaxes, providing a practical way for users to fight the spread of misinformation online.

2. METHODS

The overall methodology for this study is depicted in Figure 1. The overall methodology for this study comprises five major stages: data collection, data cleaning, data labeling, data preprocessing, and modeling.

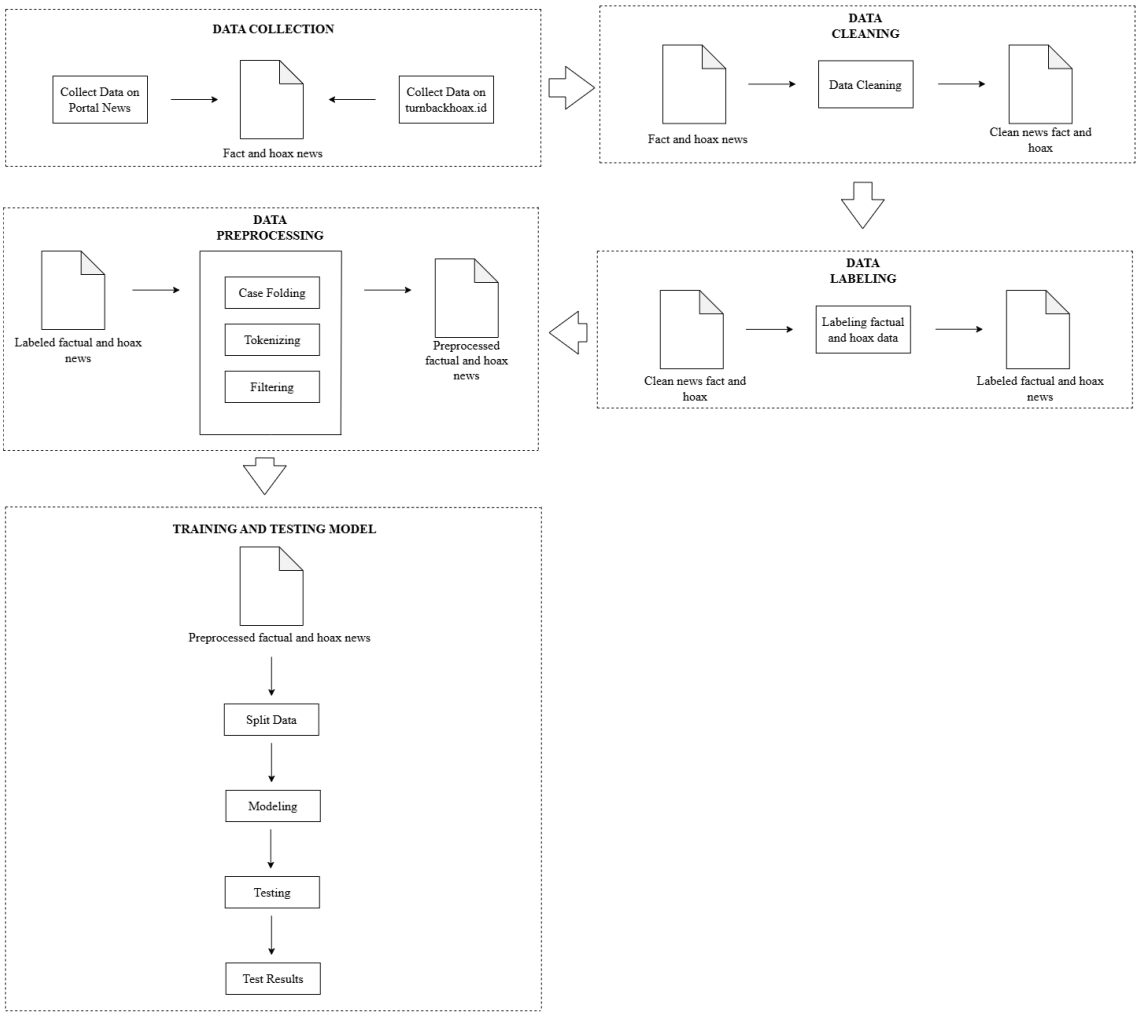

Figure 1 Overall methodology of the study

In the data collection phase of this research, the data was sourced from two distinct origins to ensure a comprehensive dataset. Firstly, genuine news articles were meticulously scraped from the online platforms of CNN Indonesia and Detik News. These two news outlets

were specifically chosen due to their established reputation and recognition as the most trusted news sources within Indonesia, a finding supported by the Reuters Institute for the Study of Journalism in their 2021 report [19]. Subsequently, the false news articles are scrapped from turnbackhoax.id, which is provided by the Ministry of Communication and Digital Affairs. The article for both factual and hoax news are gathered from September 8th until September 17th 2024. To gather relevant articles, a targeted approach using several pertinent keywords associated with President Joko Widodo was employed with specialized scraping tools. The keywords utilized in this process included "Jokowi", "Joko Widodo", "Presiden Joko Widodo", and "Presiden Jokowi". During the scraping process, various components of each news item were extracted, encompassing the title, the full textual content, and associated metadata.

Following the data acquisition, the data cleaning process was implemented with the primary objective of ensuring that the resulting dataset contained only instances that were directly relevant to the scope of this study. To achieve this, several standard and widely accepted data cleaning methodologies were applied. These included the identification and subsequent removal of any duplicate news articles that might have been collected. Additionally, for the news articles originating from turnbackhoax.id, which serves as a repository of identified false news, any extraneous tags or annotations embedded within the text were systematically removed to ensure a clean focus on the textual content. As illustrated in Table 1, the news items from turnbackhoax.id often include specific tags indicating the nature or category of the hoax, which are all used except for tag "[BENAR]" (true). Furthermore, a crucial step in the cleaning process involved the careful examination and removal of any news titles that were deemed to be irrelevant to the central topic of President Joko Widodo, ensuring that the dataset remained focused on the specific subject of interest.

Table 1 Example of tag removal for the hoax news

Hoax news titles before tag removal	**Hoax news titles after tag removal**
[SALAH] JOKOWI PERINTAHKAN KAPOLRI UNTUK PENJARAKAN GANJAR	*JOKOWI PERINTAHKAN KAPOLRI UNTUK PENJARAKAN GANJAR*
[PENIPUAN] Pemerintah Bagikan Bantuan Korban Judi Online melalui Situs Judi Online	*Pemerintah Bagikan Bantuan Korban Judi Online melalui Situs Judi Online*
[SALAH] Presiden Jokowi Ditangkap oleh Massa	*Presiden Jokowi Ditangkap oleh Massa*
[HOAX] Presiden Jokowi Membeli Kendaraan Dinas Mercedes Pullman Guard C600 dengan Harga Rp 12 M	*Presiden Jokowi Membeli Kendaraan Dinas Mercedes Pullman Guard C600 dengan Harga Rp 12 M*
[SALAH] Uang Kertas Bergambar Messi dan Timnas Argentina	*Uang Kertas Bergambar Messi dan Timnas Argentina*
[SALAH] Massa Berhasil Menduduki Istana dan Lengserkan Presiden Jokowi	*Massa Berhasil Menduduki Istana dan Lengserkan Presiden Jokowi*
[HOAX + HASUT] POLITIK LUAR NEGERI JOKOWI DI ASEAN HANYA UNTUK MENCARI INVESTASI DAN UTANG	*POLITIK LUAR NEGERI JOKOWI DI ASEAN HANYA UNTUK MENCARI INVESTASI DAN UTANG*

The labeling process followed the data cleaning stage. In this research, the labeling process was straightforward, as the source of the news served as the basis for the labels. News originating from CNN Indonesia or Detik News was labeled as factual (or positive), while news from turnbackhoax.id was labeled as hoax (or negative).

Figure 2 presents the distribution of news items across each class after the cleaning and labeling process. The dataset exhibits a significant class imbalance, approximately a one-to-ten ratio, necessitating specific handling techniques for imbalanced data, which will be detailed further in the modeling section.

The data preprocessing stage was conducted immediately before modeling to prepare the dataset for analysis. Common preprocessing techniques employed included case folding, tokenization, and word filtering, where only relevant words were selected for modeling, including the removal of stopwords. However, the stopword removal process is skipped for BiLSTM model, since it would alter the semantic meaning of the texts [20].

Following preprocessing, the modeling phase commenced. The data was split into training and testing sets, with the training data used for model development and the testing data for evaluation. This study examined several scenarios to identify the optimal model, based on a traditional model (SVM), an ensemble model (XGBoost), and a deep learning architecture (Bidirectional LSTM). For SVM and XGBoost, Term Frequency-Inverse Document Frequency (TF-IDF) was used for vectorization, as these models require document-to-vector input [21, 22]. Conversely, for BiLSTM, a trained word embedding was employed, with the embedding training occurring concurrently with the model training.

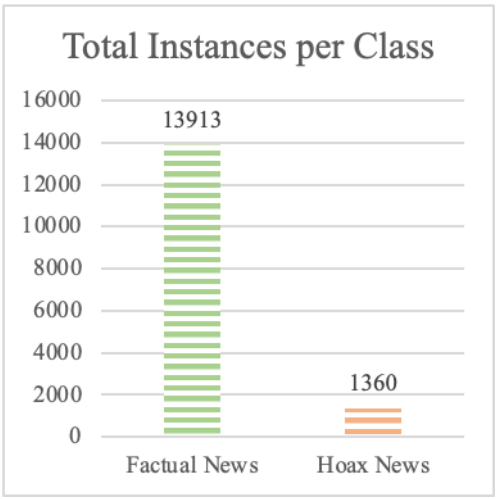

Figure 2 Total instances per class in the dataset

To address the class imbalance issue, this study implemented the Synthetic Minority Over-sampling Technique (SMOTE). Briefly, SMOTE adds new instances to the minority class by generating synthetic examples that are replications of the existing data. Utilizing a random control variable between 0 and 1 and the k-nearest neighbor algorithm, SMOTE creates new instance vectors to balance the class distribution [23]. The application of SMOTE before modeling was a key aspect of the research experiment.

This research examined several methods, both with and without the application of SMOTE for handling the imbalanced dataset. As previously mentioned, SVM, XGBoost, and BiLSTM were deployed in combination with SMOTE. Table 2 outlines the conducted experiments in this study.

Table 2 Conducted experiments

Scenario	Selected Model	Inclusion of SMOTE
Scenario 1	SVM	Without SMOTE
Scenario 2	SVM	With SMOTE
Scenario 3	XGBoost	Without SMOTE
Scenario 4	XGBoost	With SMOTE
Scenario 5	BiLSTM	Without SMOTE
Scenario 6	BiLSTM	With SMOTE

Based on Figure 2, the number of instances are extremely imbalanced, rendering evaluation with overall accuracy ineffective. Classification algorithms tend to prioritize patterns

in the majority class, leading to biased models. This can result in high overall accuracy but poor performance on minority classes, as they are often overlooked due to the strong performance on the majority class [24]. Therefore, this research employed alternative evaluation metrics such as precision, recall, and F1-score, calculated for each class [25]. Subsequently, the macro average for each metric was calculated, providing an unweighted average of the metric across all classes [26]. The results of these scenarios will determine the optimal combination and will be discussed further.

3. RESULTS AND DISCUSSION

By splitting the dataset into 80:20 ratio between training data and testing data respectively, this research obtained the concluded measure for each metric in Table 3.

Table 3 Result of each scenario

Scenario	Class	Precision	Recall	F1-score
Scenario 1 (SVM without SMOTE)	Factual news	0.96	**1.00**	**0.98**
	Hoax news	**0.92**	0.58	0.71
	Macro average	**0.94**	0.79	0.85
Scenario 2 (SVM with SMOTE)	Factual news	0.97	0.98	**0.98**
	Hoax news	0.76	0.72	0.74
	Macro average	0.87	0.85	0.86
Scenario 3 (XGBoost with out SMOTE)	Factual news	0.95	0.99	0.97
	Hoax news	0.86	0.46	0.60
	Macro average	0.91	0.73	0.79
Scenario 4 (XGBoost with SMOTE)	Factual news	0.95	0.99	0.97
	Hoax news	0.87	0.49	0.63
	Macro average	0.91	0.74	0.80
Scenario 5 (BiLSTM without SMOTE)	Factual news	0.97	0.99	**0.98**
	Hoax news	0.89	0.74	**0.81**
	Macro average	0.93	0.87	**0.90**
Scenario 6 (BiLSTM with SMOTE)	Factual news	**0.99**	0.93	0.96
	Hoax news	0.52	**0.85**	0.64
	Macro average	0.76	**0.89**	0.80

The summary of the test results from various hoax news headline classification scenarios that have been explained previously is illustrated in Table 3. The green text indicates the best measure in factual news, while the orange text reveals the best measure for hoax news and the blue text marks the best macro average. Based on the Table 3, it is revealed that the BiLSTM model without the application of SMOTE provides the best results compared to other scenarios. This is indicated by a higher macro average in the f1-score value compared to other methods and scenarios, although there are several scenarios with better measurement for precision and recall for each class.

The application of the SMOTE to SVM and XGBoost methods demonstrably led to an increase in overall performance, as indicated by the f1-score metric, although the magnitude of these improvements was relatively modest. This positive effect is clearly observable when comparing the results of Scenario 1 to Scenario 2 for SVM, and Scenario 3 to Scenario 4 for XGBoost, as presented in the study's findings. The incorporation of imbalanced data handling through the application of SMOTE did contribute to an increase in the recall rates of both models. However, this improvement in recall came with a corresponding trade-off in the form of declining precision rates. Consequently, one might initially argue that the inclusion of SMOTE into the modeling pipeline yields only marginal benefits, given the seemingly small increments in overall performance. Nevertheless, if the focus of the model is specifically directed towards

the accurate identification of hoax news – which in the context of this study is strategically positioned as the minority class due to its lower prevalence compared to factual news – then the application of SMOTE prior to the model training process proves to be an effective strategy. As evidenced by the data presented in Table 3, the performance of both the SVM and XGBoost models on the task of classifying hoax news consistently exhibits better metrics when SMOTE is taken into account during the data preprocessing stage, with the notable exception of the precision score observed in the SVM method.

Conversely, the BiLSTM model yielded a different pattern of results in response to the application of SMOTE. Unlike the SVM and XGBoost algorithms, which rely on document-to-vector representations such as TF-IDF as their input features, BiLSTM, being an architecture rooted in neural networks, utilizes word embeddings to represent the textual data. The application of SMOTE before training the BiLSTM model unexpectedly led to a decrease in its overall performance. This counterintuitive finding aligns with the results of a related study conducted by Ridho et al. [10], which demonstrated that the application of SMOTE to deep learning models, specifically citing the Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture, resulted in a significant reduction in model performance. The likely underlying reason for this performance degradation is the phenomenon of overfitting, where the model learns the characteristics of the synthetically generated oversampled data too well, hindering its ability to generalize to unseen, real-world data. This observation strongly suggests that the application of SMOTE is not universally beneficial across all types of classification models, and its utility appears to be particularly limited, or even detrimental, when employed with neural network-based models that utilize word embeddings as their input representation.

Ultimately, the BiLSTM deep learning algorithm demonstrated a higher degree of effectiveness in classifying hoax news titles compared to the SVM and XGBoost machine learning algorithms. A primary factor contributing to this superior performance is the inherent neural network architecture of the long-short term memory gates, which enables a deeper and more nuanced understanding of contextual patterns within the textual data. Furthermore, its bidirectional processing capability – where the input data is processed both from the beginning to the end and from the end to the beginning – plays a crucial role in its effectiveness. This bidirectional approach allows the model to comprehend the meaning of words not only based on the words that precede them but also on the words that follow them, as supported by the findings of Puteri [27]. This comprehensive understanding of the intricate relationships between words is of paramount importance for accurately interpreting the true meaning conveyed by a news title, especially when distinguishing between factual reporting and intentionally misleading content.

When comparing the traditional machine learning methods employed in this study, the SVM algorithm exhibited a higher level of performance than the XGBoost algorithm. This finding presents a contrast to the results reported in the research conducted by Khanam et al. [15], which indicated a marginally superior performance for XGBoost over SVM. This discrepancy in the observed performance could be attributed to the inherent differences in the linguistic characteristics and the underlying structure of the datasets utilized in the two studies. Specifically, Khanam's research employed a dataset comprised of English-language text, whereas the current research focused on an Indonesian-language dataset. The Indonesian language possesses distinct grammatical structures, a unique vocabulary, and different levels of linguistic complexity compared to English, factors that can differentially influence the effectiveness of various classification algorithms during the process of learning and categorization.

Based on the data presented in Table 3, across all the evaluated experimental scenarios where no oversampling techniques were applied, the precision metric consistently remained higher than the recall metric. This observation indicates a notable strength in the models' ability to accurately identify a news title as a hoax when they make such a classification, thereby minimizing the occurrence of false positives on legitimate news content. However, the

corresponding lower recall values suggest that while the models exhibit reliability in their positive predictions, they are likely failing to identify a considerable proportion of the actual hoax news items present within the dataset. In the context of minimizing the harm of hoaxes, the model is expected to capture as many of the actual hoaxes as possible to prevent the spread of misinformation. Therefore, recall metric is more considerable than precision for hoax detection, although it comes with a more lenient model by flagging even factual news as hoax. On the other hand, when oversampling techniques were applied, all models denotes increases in recall score for hoax news. However, as expected, corresponding decreases in precision as a trade-off are evident for each model, with BiLSTM model's precision exhibited the steepest decline but compromised by highest recall score. This outcome highlights the BiLSTM model's potential to identify a larger proportion of actual hoax news, especially when the training data is balanced.

The misclassification found in this study reveals no systematic pattern. However, the models demonstrate a small propensity to misclassify an actual news when one or multiple negative sentiment terms are present in the respective title. For instance, the term *diusir* (kicked out) frequently appears in hoax-labelled news compared to the actual news. When an actual news title contain this term, such as "*NasDem Undang Jokowi di Kongres III: Kita 10 Tahun Setia Meski Diusir"* is prompted into model, it is consequently misclassified as hoax news. Correspondingly, this research indicates that hoax news titles are predominately biased to negative sentiment. Therefore, the neutral-toned news are generally categorized as actual news by the model. For example, the hoax news with title "*Jokowi dan Prabowo ke Qatar Dukung Timnas*" (Jokowi and Prabowo goes to Qatar to Support the national team) is misclassified as actual news by the model, since the title shows no polarized sentiment. This finding suggests a considerable direction for future works: a title-only classification model may lack the necessary contextual depth to accurately capture the subject of the news, since the content of the news is often underrepresented by the limited-word title.

4. CONCLUSIONS

This research investigated the efficacy of various machine learning (SVM, XGBoost) and deep learning (BiLSTM) algorithms for classifying hoax news by focusing on news titles, aiming to streamline detection and minimize traffic to hoax-spreading websites. The methodology involved collecting and preprocessing a dataset of factual news (from CNN Indonesia and Detik News) and hoax news (from turnbackhoax.id), followed by training and evaluating the aforementioned models, both with and without the application of the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance.

The findings indicate that the BiLSTM deep learning algorithm generally outperformed the traditional machine learning approaches in this Indonesian language context, likely due to its superior ability to capture contextual patterns in text. However, the application of SMOTE negatively impacted BiLSTM's performance, suggesting potential overfitting on synthetic data when using word embeddings. Furthermore, the study's comparison of SVM and XGBoost performance with prior research on English datasets underscores the importance of considering language-specific characteristics in hoax detection. Ultimately, this title-based classification approach, particularly with deep learning models, shows promise as a more efficient method for automated hoax detection in Indonesia, although careful consideration of data balancing techniques and linguistic nuances is crucial for optimal results.

ACKNOWLEDGEMENTS

We sincerely thank our colleagues and senior members of the Information Technology Study Program at Udayana University for their invaluable guidance and support throughout this research. Their expertise and insights were crucial in shaping our work and helping us overcome challenges.

REFERENCES

[1] R. N. Rahayu, "Vaksin covid 19 di Indonesia: analisis berita hoax," *Jurnal Ekonomi, Sosial & Humaniora*, vol. 2, no. 07, pp. 39-49, 2021.

[2] C. Juditha, "Hoax communication interactivity in social media and anticipation (Interaksi komunikasi hoax di media sosial serta antisipasinya)," *Pekommas*, vol. 3, no. 1, p. 261723, 2018.

[3] S. Soleman, "Pemanfaatan Metode Klasifikasi Naïve Bayes Untuk Pendeteksi Berita Hoax Pada Artikel Berbahasa Indonesia," *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, vol. 7, no. 2, pp. 83-93, 2021.

[4] R. R. Sani *et al*., "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *Jurnal Masyarakat Informatika*, vol. 13, no. 2, pp. 85-98, 2022.

[5] E. Utami *et al*., "Covid-19 Hoax Detection Using KNN in Jaccard Space," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, pp. 255-264, 2021.

[6] S. Chitti *et al*., "Identifying Hoaxes in Fake Spotter using XG Boost Machine Learning based Classification Method," in *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, Dec. 2024, pp. 990-996.

[7] A. Mohapatra, N. Thota, and P. Prakasam, "Fake news detection and classification using hybrid BiLSTM and self-attention model," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18503-18519, 2022.

[8] A. D. Cahyani and A. K. Ramdani, "Hoax Detection of Covid-19 News on Social Media using Convolutional Neural Network (CNN) and Support Vector Machine (SVM)," *International Journal on Information and Communication Technology (IJoICT)*, vol. 9, no. 2, pp. 177-185, 2023.

[9] C. Agrawal, A. Pandey, and S. Goyal, "Multimodal fake news detection using hyperparameter-tuned BERT and ResNet110," *International Journal of Advanced Technology and Engineering Exploration*, vol. 11, no. 114, p. 759, 2024.

[10] M. Y. Ridho and E. Yulianti, "From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News," *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, vol. 10, no. 3, pp. 544–555, 2024, doi: 10.26555/jiteki.v10i3.29450.

[11] L. H. Suadaa, I. Santoso, and A. T. B. Panjaitan, "Transfer learning of pre-trained transformers for covid-19 hoax detection in indonesian language," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, pp. 317-326, 2021.

[12] P. W. Cahyo and U. S. Aesyi, "Perbandingan LSTM dengan Support Vector Machine dan Multinomial Na ve Bayes pada Klasifikasi Kategori Hoax," *Jurnal Transformatika*, vol. 20, no. 2, p. 23, 2023, doi: 10.26623/transformatika.v20i2.5880.

[13] R. K. Putri and M. Athoillah, "SUPPORT VECTOR MACHINE UNTUK IDENTIFIKASI BERITA HOAX TERKAIT VIRUS CORONA (COVID-19)," vol. 6, no. 3, 2021.

[14] I. A. Ropikoh, R. Abdulhakim, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Berita Hoax Covid-19," *Journal of Applied Informatics and Computing (JAIC)*, vol. 5, no. 1, 2021, Accessed: Apr. 30, 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," in *IOP conference series: materials science and engineering*, vol. 1099, no. 1, Mar. 2021, p. 012040.

[16] J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 5, p. 052081, 2021, doi: 10.1088/1757-899x/1098/5/052081.

[17] H. B. Aji and E. B. Setiawan, "'Detecting hoax content on social media using bi-LSTM and RNN," *Building Informat., Technol. Sci*, vol. 5, no. 1, pp. 114-125, 2023.

[18] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods," *ICT Express*, vol. 8, no. 3, pp. 396–408, 2022, doi: 10.1016/j.icte.2021.10.003.

[19] Reuters Institute for the Study of Journalism, "Digital News Report 2021," University of Oxford, Oxford, UK, 2021. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (Accessed: Apr. 30, 2025)

[20] A. Santosa, I. Purnamasari, and R. Mayasari, "Pengaruh Stopword Removal dan Stemming Terhadap Performa Klasifikasi Teks Komentar Kebijakan New Normal Menggunakan Algoritma LSTM," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 6, no. 1, pp. 81-93, 2022.

[21] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, Oct. 2020, pp. 1-6.

[22] Zhang Qi, "The text classification of theft crime based on TF-IDF and XGBoost model," in *2020 IEEE International conference on artificial intelligence and computer applications (ICAICA)*, 2020, pp. 1241-1246.

[23] E. Sutoyo and M. Asri Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, vol. 6, no. 3, pp. 379–385, 2020.

[24] R. D. Firtriani, H. Yasin, and Tarno, "Penanganan Klasifikasi Kelas Data Tidak Seimbang dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," vol. 10, no. 1, pp. 11–20, 2021.

[25] Y. Ansori and K. F. H. Holle, "Perbandingan Metode Machine Learning dalam Analisis Sentimen Twitter," *Jurnal Sistem Dan Teknologi Informasi (JustIN)*, vol. 10, no. 4, p. 429, 2022, doi: 10.26418/justin.v10i4.51784.

[26] A. R. Hanum *et al*., "Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, pp. 537-546, 2024.

[27] D. I. Puteri, "Implementasi Long Short Term Memory (LSTM) dan Bidirectional Long Short Term Memory (BiLSTM) Dalam Prediksi Harga Saham Syariah," *Euler : Jurnal Ilmiah Matematika, Sains Dan Teknologi*, vol. 11, no. 1, pp. 35–43, 2023, doi: 10.34312/euler.v11i1.19791.