Jurnal Sistem Informasi Bisnis 02(2025)
Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587
On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/72212

Improving Fake News Detection Accuracy with Lexicon-based Approach and LSTM through Text Preprocessing

Chamdan Mashuri[a,*], Edwin Hari Agus Prastyo[b], Fajar Rohman Hariri[c]

[a] Information System, Faculty of Information Technology, Hasyim Asy'ari University Tebuireng Jombang
[b] Informatics Engineering, Faculty of Information Technology, Hasyim Asy'ari University Tebuireng Jombang
[c] Informatics, Faculty of Science and Technology, State Islamic University of Malang

Submitted: April 1st, 2025; Revised: June 22nd, 2025;
Accepted: June 23rd, 2025; Available Online: June 29th, 2025
DOI: 10.14710/vol15iss2pp285-292

Abstract

Fake news detection has become a critical issue in the digital era, especially with the rapid growth of social media and online platforms. In Indonesia, effective detection of hoaxes plays a vital role in preventing social unrest, maintaining public trust, and promoting informed decision-making. This research aims to enhance the accuracy of detecting fake news in Indonesian by developing a model using lexicon-based and Long Short-Term Memory (LSTM) approaches. The study integrates sentiment analysis with lexicon-based scoring to identify key features in news articles, while LSTM is employed to analyze sequential patterns in the data. The methods were tested on a dataset consisting of both hoax and non-hoax news collected from reliable sources. The results indicate that the hybrid model significantly improves the detection accuracy, achieving an impressive accuracy rate of 99%. This research demonstrates the potential of combining lexicon-based and LSTM approaches to overcome challenges in detecting fake news, especially in low-resource languages like Indonesian. The findings contribute to advancing the development of reliable and efficient systems for combating misinformation in the digital age.

Keywords: Fake News Detection; Lexicon-Based; LSTM; accuracy; Algorithm.

1. Introduction

Fake news is a growing problem in this digital era, especially with the rapid use of social media and internet-based platforms. The Ministry of Communication and Information Technology (Kominfo) noted that as of May 2023, as many as 11,642 hoaxes had been identified, with the categories of health (2,287), government (2,111), fraud (1,938), and politics (1,373) being the most common. In addition, the survey shows that between 30% and almost 60% of internet users in Indonesia are exposed to hoaxes when accessing and communicating online, while only 21% to 36% are able to recognise these fake news. The spread of fake news can have a wide range of negative impacts, from creating social unrest and influencing public opinion to damaging political and economic stability (Kominfo, n.d.). This figure shows that the spread of hoaxes is still rampant and needs to be taken seriously. Hoaxes can cause various negative impacts, such as public distrust of the government and mass media, political polarisation, social unrest, and economic and security losses (P *et al.*, 2023). Therefore, this research focuses on developing a hoax news detection model that can help identify and counteract the spread of false information, thus making a positive contribution to society in facing information challenges in the digital era.

Previous research has shown various approaches in detecting hoax news. For example, research by Zaman *et al.* developed a hoax news detection system in Indonesia using reader feedback and Naïve Bayes algorithm (Hutama and Suhartono, 2022) . Hutama and Suhartono (Sudrajat *et al.*, 2022) successfully classified Indonesian hoax news with multilingual transformer models and BERTopic, which showed that these methods were able to outperform basic models in classifying hoax news in low-resource languages, such as Indonesian. Research by Prasetya and Ferdiansyah (Prasetya and Ferdiansyah, 2022) highlights the importance of larger and balanced datasets to improve the accuracy and generalisability of hoax news detection models. This reference shows that the use of the right dataset can affect the performance of the detection model.

In addition, fake news detection can be improved by incorporating sentiment analysis as an important feature (Balshetwar *et al.*, 2023). Lexicon-based methods are used to develop key feature words with sentiment scores using lexicon-based algorithms. The utilisation of Long Term Memory models (LSTM) has

*) Corresponding author: chamdanmashuri@unhasy.ac.id

proven to be a significant advancement in combating the spread of misinformation. LSTM, as a form of deep learning algorithm (Baidawi, 2021) , is able to identify hoax news by analysing patterns and characteristics of text data in news articles. The study by Gohan (Ade Gohan *et al.*, 2021) highlights the importance of a participatory approach in responding to hoax news, while the study by Pardede and Ibrahim (Pardede and Ibrahim, 2020) implements LSTM for English-language hoax news identification.

These studies show methodological gaps, especially in the use of small and unbalanced datasets and limited research using Lexicon-based and LSTM approaches for hoax news detection in Indonesian. This research aims to develop and test a Lexicon-based hoax news detection model and LSTM that can overcome the problem of unbalanced datasets and improve the accuracy of hoax news detection in Indonesian. It is hoped that this research can make a significant contribution to the development of more effective and reliable hoax news detection methods, and help the Indonesian people in counteracting the spread of false information that can have a negative impact on social and political life.

The use of Lexicon-based and LSTM methods in hoax news detection is important because it can increase the effectiveness in identifying misinformation. Lexicon-based approaches enable sentiment analysis that involves developing key words with sentiment scores using lexicon-based algorithms (Al-Shabi, 2020) . Meanwhile, the utilisation of the Long-Term Memory model (LSTM) has brought significant progress in combating the spread of misinformation, especially in identifying hoax news by analysing patterns and characteristics of text data. In the Indonesian context, further research is needed to test the reliability and effectiveness of this approach in detecting hoax news in various environments, including in rural areas and among internet users with low levels of digital literacy. Thus, the use of Lexicon-based and LSTM methods can help improve the quality and accuracy of hoax news detection systems, especially in the face of the rapid spread of fake news through social media and online platforms (Tama and Sibaroni, 2023). Research by Pulungan (Pulungan, 2022) discusses the perspective of the Angkola Muslim community in responding to hoaxes through cognitive mental process learning, highlighting the importance of cognitive understanding in responding to hoax news. In addition, the study by Yunanto (Yunanto *et al.*, 2021a) summarises fake news detection strategies with four different approaches, including source credibility, which can provide insight into factors that affect detection accuracy. The study by Khanifah and Fauzi Khanifah (Khanifah and Fauzi, 2022) highlights the impact of hoax news on the implementation of health protocols by communities, while the study by Saraswati (Saraswati *et al.*, 2022) addresses the importance of anti-hoax skills training for students. Thus, more inclusive and diverse research can help understand and address population gaps in hoax news detection in Indonesia.Based on the identification of research gaps above, this study is expected to contribute to the knowledge and practice of hoax news detection in Indonesia

This research focuses on lexicon-based and LSTM approaches in hoax news detection. These two approaches were chosen due to their proven effectiveness in previous studies, showing satisfactory results in recognising hoax news patterns in the context of complex digital information. The complementarity between lexicon-based and LSTM is also a consideration, where lexicon-based uses a dictionary of key words related to hoax news, while LSTM utilises neural networks to recognise complex and unstructured patterns. By combining these two approaches, it is expected to utilise the strengths of each approach and overcome their weaknesses, resulting in higher detection accuracy. In addition, the adaptability of both lexicon-based and LSTM for the Indonesian language is an additional reason for choosing this approach, which is expected to make a significant contribution in improving the accuracy of hoax news detection on Indonesian news sites.

This research will focus on developing a hoax news detection model using lexicon-based and LSTM approaches. Lexion-based method is used to detect fake news based on a dictionary of words that indicate hoaxes, generally hoax news is made with provocative and slanderous titles, with language that can convince readers (Zhang *et al.*, 2020) , evaluation of the effectiveness of the developed model using hoax and non-hoax news datasets using the LSTM method, as well as analysis of factors that affect model accuracy. The contribution of the results of this research is highly expected in developing the field of hoax news detection. First, the results are expected to improve the accuracy of hoax news detection on Indonesian news sites through the developed model. Second, this research will enrich the knowledge of hoax news detection by providing new insights into the optimisation of lexicon-based and LSTM approaches. Third, the results of this research are expected to provide assistance for developers and practitioners in building a more effective hoax news detection system, so as to better maintain information integrity and public security.

Thus, through a technological approach as previously researched, for this research to raise a title Improving the Accuracy of Hoax News Detection on Indonesian News Sites with Lexicon-Based and LSTM Approaches, it is expected to detect hoax news on Indonesian online news sites with perfect results. This study represents significant progress in reducing the adverse impact of the proliferation of false information on public welfare and community well-being. The spread of false information has the capacity to trigger societal fragmentation, social unrest, and jeopardise public safety. Conventional methodologies such as manual authentication have become inadequate given the increasing volume of false information. Therefore, this novel research with lexicon-based approach and LSTM shows promising potential in improving the accuracy of automatically identifying false information, an important measure in improving the accuracy in hoax detection.

2. Theory

2.1 Sentiment Analysis Based on Lexion

Lexicon-based methodologies are commonly used in sentiment analysis models in research studies to categorise sentiments by using a lexicon or corpus containing weighted language words as linguistic resources (Ni Made Ayu Juli Astari *et al.*, 2020) The main purpose of building such a lexicon is to reduce instances of misclassification that may occur when the test dataset contains ambiguous sentiment scores (Kurniawan and Mustikasari, 2021) The use of lexicon-based approaches has shown potential in improving the effectiveness of sentiment classification tasks (Yunanto *et al.*, 2021) Typically, the feelings considered include positive and negative polarities. The process of creating positive and negative lexicon word dictionaries involves segmenting the tweet text into word fragments to determine positive and negative sentiment scores.

The calculation of lexicon-based scores is done through equations 1 and 2, where the sum of scores above a certain threshold th is classified as positive, and below t as negative.

$$\text{if } w\text{Score} > t \Rightarrow \text{positive (hoax)} \tag{1}$$

$$\text{if } w\text{Score} < t \Rightarrow \text{negative (fact)} \tag{2}$$

2.2 LSTM (Long Short-Term Memory)

Long Term Short Term Memory (LSTM) networks[20] are a special type of recurrent neural network (RNN)[21] developed to overcome the vanishing gradient problem encountered in traditional RNNs[22] . LSTM networks excel in handling gap lengths with a certain degree of indifference, setting them apart from other RNNs, hidden Markov models, and alternative sequence learning approaches. The key strength of the LSTM lies in its ability to retain short-term memory over many stages of time, hence the name "long-term short-term memory." These models are widely used in various applications such as classification, data processing, and predictive analysis involving time series data, including domains such as handwriting recognition, speech recognition, machine translation, speech activity detection, robot control, video games, and healthcare.

A brief representation of the equations governing the forward propagation process in a Long Short-Term Memory (LSTM) cell featuring forgetting gates is:

$$f_t = g(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = g(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = g(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = g(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$
$$h_t = o_t \cdot \tanh(c_t)$$

Let the superscripts d and h refer to the number of input features and the number of hidden units, respectively.

- $x_t \in \mathbb{R}^d$: input vector to the unit
- $f_t \in (0,1)^h$: forget gate's activation vector
- $i_t \in (0,1)^h$: input/update gate's activation vector
- $o_t \in (0,1)^h$: output gate's activation vector
- $h_t \in (-1,1)^h$: hidden state vector, also known as the output vector of the LSTM unit
- $\tilde{c}_t \in (-1,1)^h$: cell input activation vector
- $c_t \in \mathbb{R}^h$: cell state vector
- $W_* \in \mathbb{R}^{h\times d}$, $U_* \in \mathbb{R}^{h\times h}$, and $b_* \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training.

3. Methods

This research methodology aims to improve the accuracy of fake news identification by using lexicon-based strategies and Long-Term Memory (LSTM). Lexicon-based strategies are used to detect prominent words and phrases that are often found in fake news. In contrast, LSTM models are used to capture sequential patterns in news articles and improve the forecasting ability of the detection system. This investigation includes several important phases: data acquisition, data pre-processing, lexicon-based model development, LSTM model training, and model performance evaluation. During the data collection phase, a series of news outlets were recognised and combined to build a comprehensive data set. Data pre-processing involves text enhancement, tokenisation, and the creation of a customised lexicon for fake news. The creation of a lexicon-based model entails the recognition and assimilation of words and phrases that commonly manifest in fake news. Following this, the LSTM model is trained using the pre-processed data to identify trends and structures in the news content. Ultimately, the efficacy of the model is measured through metrics such as accuracy, precision, and recall to ascertain the feasibility of the proposed methodology.

Data collection was conducted from the fact and fake news provider website, Turnbackhoax.id, managed by Masyarakat Anti Hoax Indonesia (MAFINDO) since November 2016. The content comes from Forum Anti Fitnah Hasut and Hoax (FAFHH), with each news item marked as fact or hoax. The data consisted of 2160 hoax news from the site and 2152 non-hoax news from detik.com. The collection also involved the use of 23 hoax keywords in the lexicon and 758 stopwords in Indonesian

After collection, the dataset was divided with an allocation of 70% for train data and 30% for test data. The technique used to collect the data was web scraping, which allows automatic retrieval of data from the mentioned sources.

3.1 Hoax and Non Hoax Labelling Process

Labelling is done for the purpose of annotating news articles that have previously been collected and stored in the databerita.csv file. An illustration of the data labelling procedure is presented in Table 1.

Tabel 1.Illustration of data labelling procedure

Text	Label
[FALSE] Formula E Race's Peak Event Lacks Crowds	1
[FALSE] "HRS reportedly killed by camel at camel race in Saudi Arabia"	1
[FALSE] Formula E Not Broadcasted by National TV	1
[FALSE]: Formula E Jakarta sets world record for least crowded race event	1
Strategic foresight for presidential candidates	0
Trade Minister Zulhas Explains the Effect of Geopolitics on Indonesian Trade	0
When will the Prabowo-Gibran TKN be announced? This is Gerindra's answer	0

An illustration of the data labelling procedure shows example texts and their labels. Texts that contain false information, such as "Formula E Race Summit Event Deserted" and "Formula E Not Broadcasted by National TV," are labelled 1, indicating that the news is a hoax. Conversely, texts that contain correct information, such as "Strategic Foresight for Presidential Candidates" and "Minister of Trade Zulhas Explains Geopolitical Influence on Indonesian Trade," are labelled 0, indicating that the news is a fact. This labelling process is important to classify and identify news based on its truthfulness, facilitating further analysis related to the accuracy of the information conveyed in the news context.

3.2 Data Preprocessing

This function aims to process the text with several preprocessing steps. Firstly, the text is converted into lowercase letters to ensure uniformity in the data. Then, numeric numbers are removed from the text, followed by the removal of non-word characters such as punctuation marks. Finally, additional blank spaces are removed to clean up the text as a whole.

Table 2. Data preprocessing results

News Text Before Preprocessing	News Text After Preprocessing
"At 10.00 am, there was a major fire in a residential area. 20 houses were reportedly burnt down due to electrical short circuit."	"at wib there was a large fire in a residential area, a house unit was reportedly burned due to an electrical short circuit"
"In a meeting held yesterday, it was agreed to raise the price of petrol from next week."	"In a meeting held yesterday, it was agreed to raise the price of fuel oil starting next week"
"Based on the latest data, the number of Covid-19 cases in the city has reached 100,000 with 20,000 of them still under treatment."	"Based on the latest data, the number of covid cases in this city has reached thousands with thousands of them still under treatment"
"A new study has found that drinking coffee regularly may reduce the risk of heart disease."	"a new study finds that drinking coffee regularly may reduce the risk of heart disease"

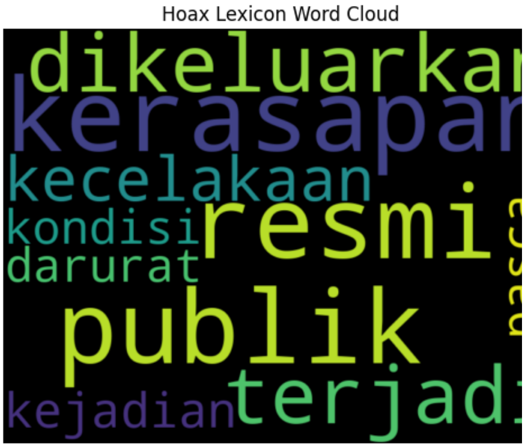
Table 2: Data Preprocessing Results shows the text processing before and after the preprocessing stage. Firstly, the text is converted into lowercase letters to ensure uniformity in the data. Next, numeric numbers are removed from the text, followed by the removal of non-word characters such as punctuation marks. Finally, additional blank spaces are removed to clean up the text as a whole. For example, the original text containing a specific time and number of units ("At 10.00 am, there was a major fire in a residential area. 20 housing units were reported burnt down due to electrical short circuit.") has been converted into a simpler and easier to process text ("at 10:00 am there was a major fire in a residential area housing units were reported burnt down due to electrical short circuit") after going through the preprocessing process. This process enables more accurate and efficient analysis of text content in studies or other text-based applications.

4. Results and Discussions

In this chapter, we will present the results of experiments to test the effectiveness of lexicon-based and Long Short-Term Memory (LSTM) approaches in detecting fake news. The results obtained, including accuracy, precision, and recall, will be analysed and compared with other models. Next, we will discuss the factors that influence model performance, the impact of text preprocessing, and the combined contribution of the two approaches in improving accuracy. This chapter will also compare the research results with previous studies and identify potential for further development.

4.1. Lexicon-Based Results

The results show that creating a lexicon with words/phrases associated with hoaxes such as clickbait, sensational, and provocative has a significant impact on hoax news detection. The first step is to match the words/phrases in the news with the lexicon that has been created. This process is done by looking for matches between the words/phrases in the news text and the entries in the lexicon. Next, a score is given based on the number and type of words/phrases found in the news. This score gives an indication of how likely the news is a hoax based on the words/phrases identified in the lexicon. Thus, creating a lexicon and scoring based on hoax-related words/phrases is an important step in improving the accuracy of hoax news detection using a lexicon-based approach.



Word lexicon-based

Figure 2 displays a word cloud containing words related to hoax news, obtained from a lexicon-based analysis of hoax-labeled news texts. The visualization highlights frequently occurring terms that are often associated with misleading or false information in Indonesian online media. Prominent words such as "dikeluarkan" (released), "resmi" (official), "publik" (public), "terjadi" (happened), and "kerusakan" (damage) suggest that hoaxes often mimic formal or authoritative language to appear credible. Terms like "darurat" (emergency), "kondisi" (condition), and "kecelakaan" (accident) indicate that many hoax articles exploit emotionally charged or urgent situations to provoke public reaction. This figure demonstrates how hoax content tends to use specific lexical patterns to manipulate perception and spread misinformation effectively.The words that appear most frequently in the word cloud are shown in a larger size. Words such as "issued", "accident", "condition", "emergency", "public", "incident", and "happened" are often used in hoax news to describe dangerous or threatening situations. This word cloud can help identify hoaxes; if a story contains many of the words seen in the word cloud, then it is more likely to be a hoax.

4.2. LSTM Result

In the LSTM (Long Short-Term Memory) model, there are several parameters that have an important role in the training and prediction process. These parameters are addressed in table 3.

Table 3. LSTM parameters

	Description
max_features	The maximum number of features that the Tokenizer will process on the text before padding.
max_len	The maximum length of the token sequence after tokenisation and padding.
embedding_dim	The embedding dimension used in the Embedding layer in the LSTM model.
epochs	Number of iterations when training the LSTM model.
batch_size	The number of data samples used in one iteration when training the model.
threshold	The threshold to determine the prediction (Hoax or Non-Hoax) from the probability obtained from the model. This value can be customised.
loss_function	Loss function used when training the LSTM model.
optimiser	The optimisation algorithm used when training the LSTM model.
dropout	The dropout value used in the SpatialDropout1D layer in the LSTM model.
recurrent_dropout	The dropout value used in the LSTM layer in the LSTM model.

Table 3, LSTM Parameters, presents various parameters that are relevant in the configuration and training of the LSTM model for text classification, specifically in identifying fake news. Parameters such as `max_features` (maximum number of features processed by the Tokenizer), `max_len` (maximum length of the token sequence after tokenisation and padding), `embedding_dim` (embedding dimension in the Embedding layer), `epochs` (number of iterations during training), `batch_size` (number of data samples in one iteration), etc., have a critical role in determining the performance and accuracy of the model. The right choice of each of these parameters can affect how well the LSTM model can classify news texts between hoaxes and non-hoaxes, with `threshold` (threshold for prediction), `loss_function` (loss function), `optimizer` (optimisation algorithm), `dropout`, and `recurrent_dropout` (dropout value in the LSTM layer) also being determining factors in effective model building and training.

The following table 4 is a test table that displays the values of the variables used in the LSTM model for testing and performance evaluation.

Table 4. LSTM variable value test

max_features	*Max len*	*Embedding dim*	*epochs*	*batch_size*	*threshold*	*dropout*	*Recurrent dropout*
3000	80	100	5	32	0.4	0.2	0.2
5000	100	128	5	64	0.5	0.4	0.2
5000	150	150	10	128	0.6	0.4	0.4
6000	100	200	8	256	0.7	0.5	0.5
7000	130	250	6	128	0.8	0.6	0.6

Table 4, which is the LSTM Variable Value Test, displays the experimental results with various parameters tested on the LSTM (Long Short-Term Memory) model. The tested variables include the maximum number of features (max_features), maximum length (max_len), embedding dimension, number of epochs, batch size, threshold, regular dropout, and recurrent dropout. This experiment aims to evaluate the performance of the LSTM model in text classification, where parameter settings such as feature size, text length, and dropout can affect the accuracy and overall performance of the model in identifying hoax news texts. One configuration that stands out from the table is number 3, with parameters that can be considered optimal in the context of performing best for the specific text classification task.

The following table 5 is a list of Optimiser mudules, Optimiser is a key parameter in the LSTM model training process that has a significant influence on model performance and accuracy. The data in this table provides an overview of the effectiveness of various optimisers in improving the quality of fake news detection predictions.

Table 5. Optimiser test

No.	*Optimiser* Name	*Model.compile* code
1	Adam	*model.compile(loss='binary_crossentropy', optimiser='adam', metrics=['accuracy'])*
2	SGD (Stochastic Gradient Descent)	*model.compile(loss='binary_crossentropy', optimiser=SGD(), metrics=['accuracy'])*
3	RMSprop	*model.compile(loss='binary_crossentropy', optimiser=RMSprop(), metrics=['accuracy'])*
4	Adagrad	*model.compile(loss='binary_crossentropy', optimiser=Adagrad(), metrics=['accuracy'])*
5	Adadelta	*model.compile(loss='binary_crossentropy', optimiser=Adadelta(), metrics=['accuracy'])*
6	Nadam	*model.compile(loss='binary_crossentropy', optimiser=Nadam(), metrics=['accuracy'])*

Table 5, Optimiser Tests, compares several commonly used optimisers in neural network model training. Each optimiser is indicated by its name and code for use in the `model.compile` function of the neural network model building process. Optimisers included in this table include Adam, SGD (Stochastic Gradient Descent), RMSprop, Adagrad, Adadelta, and Nadam. Each optimiser has different characteristics and optimisation methods, such as convergence speed, learning rate handling, and adaptation to different gradients. Choosing the right optimiser can affect the performance and results of the trained model, depending on the complexity of the dataset and the model architecture used.

4.4. Results of comparing the performance of the combination of Lexicon-Based and LSTM with each method and Test Variable parameters and Optimiser method

a. Parameter Variable Test Results

The results of testing the parameter values in LSTM which resulted in the 2nd test with labels 0 (fact) and 1 (hoax) resulted in the highest accuracy of 0.997 with a processing time of 267 seconds with the 2nd test parameter from this table showing the best results in terms of hoax news detection accuracy, test results Variable parameters in table 6.

Table 6. Optimiser test

Test to	Label	Time (s)	precision	Recalll	Fi-score	accuracy
1	0	265	1.00	1.00	1.00	0.992
	1	265	1.00	1.00	1.00	0.992
2	0	267	1.00	1.00	1.00	0.997
	1	267	1.00	1.00	1.00	0.997
3	0	512	0.99	0.99	0.99	0.991
	1	512	0.99	0.99	0.99	0.991
4	0	266	0.99	0.99	0.99	0.990
	1	266	0.99	0.99	0.99	0.990
5	0	446	1.00	1.00	1.00	0.995
	1	446	1.00	1.00	1.00	0.995

Table 6. Optimizer test, In the 2nd test, the parameters used are `max_features` of 5000, `max_len` of 100, `embedding_dim` of 128, `epochs` of 5, `batch_size` of 64, `threshold` of 0.5, `dropout` of 0.4, and `recurrent_dropout` of 0.2. The use of this parameter combination resulted in higher accuracy in predicting fake news compared to the parameter settings in the other tests. This shows that the combination of parameter values in the 2nd test is more effective in improving the quality of hoax news detection prediction using the LSTM method.

b. Test Results of Optimiser method

After conducting a series of parameter tests on the LSTM method for hoax news detection, the best results were found in the 2nd parameter test with the highest accuracy. This parameter test produces a very good accuracy value in predicting fake news. Furthermore, to strengthen these results, an optimiser test was conducted using the existing modules in LSTM.

Table 7. Test run of Optimiser Test

Optimiser	*Label*	*precision*	*Recalll*	*Fi-score*	*accuracy*
Adam	0	1.00	1.00	1.00	0.997
	1	1.00	1.00	1.00	0.997
SGD (Stochastic Gradient Descent)	0	0.50	0.99	0.66	0.507
	1	0.77	0.05	0.09	0.507
RMSprop	0	1.00	0.99	0.99	0.994
	1	0.99	1.00	0.99	0.994
Adagrad	0	0.50	0.89	0.64	0.511
	1	0.58	0.14	0.23	0.511
Adadelta	0	0.49	0.75	0.59	0.491
	1	0.50	0.24	0.33	0.491
Nadam	0	1.00	0.99	0.99	0.994
	1	0.99	1.00	0.99	0.994

Table 7, Optimizer Test Test, The results of the optimizer test with Adam's method show outstanding results, with precision, recall, F1-score, and accuracy values that are close to perfect with a result of 0.997. This shows that the use of Adam's optimiser significantly improves the prediction quality of fake news detection using the LSTM method.

4.4. Hoax Detection Accuracy Evaluation

Comparative table with previous research with the research we have done and the dataset we created called FakeFact.ID (Indonesia Fake News).

Table 7. Accuracy Evaluation

Ref	*Language*	*SA method*	*Detecion Method*	*Data Set*	*Performance*
Varol *et al.* (2017)[3]	English	Lexicon-based	KNN	Ad-hoc from Twitter	Acc = 0.97
Ajao *et al.* (2019)[4]	English	Lexicon-based	SVM	Rumours	Acc = 0.86
Anoop *et al.* (2020) [5]	English	Lexicon-based	Naive Bayes	HWB	Acc = 0.790
			KNN		Acc = 0.925
			SVM		Acc = 0.900
			Random Forests		Acc = 0.840
			Decision Tree		Acc = 0.940
			There is Boost		Acc = 0.965
			CNN		Acc = 0.910
			LSTM		Acc = 0.920
Our research	Indonesia	Lexicon-based	LSTM	Our Dataset	Acc = 0.99

4.5. Data Analysis and Interpretation of Results:

Data analysis results show that lexicon and LSTM parameters have a significant impact on hoax news detection accuracy. Parameters such as `max_features`, `max_len`, `embedding_dim`, `epochs`, `batch_size`, `threshold`, `dropout`, and `recurrent_dropout` have an important role in improving the accuracy of fake news detection using lexicon-based and LSTM approaches.

`Max_features` and `max_len` specify the number of unique words and the length of the processed text, `embedding_dim` sets the word representation dimension, `epochs` sets the training iterations, `batch_size` sets the number of samples in one iteration, `threshold` is the prediction threshold, `dropout` and `recurrent_dropout` set the overfitting.

By optimising these parameters, we can improve the accuracy of hoax news detection. In the experiment, the best parameters are `max_features=5000`, `max_len=100`, `embedding_dim=128`, `epochs=5`, `batch_size=64`, `threshold=0.5`, `dropout=0.4`, and `recurrent_dropout=0.2`, resulting in optimal accuracy.

Results from experimental studies on the hybrid model show excellent performance, with accuracy rates and other performance metrics reaching around 99%. This model can be an effective tool in combating the spread of false information that threatens society.

The combination of Lexicon-Based and LSTM approaches improves hoax detection accuracy through a structured approach. The approach involves a lexicon-based filter to identify hoax news using a dictionary of keywords. A pre-trained LSTM predicts the likelihood of news as a post-lexicon filtering of hoaxes. The LSTM prediction results and the lexicon score are combined to obtain a final score, which is used to categorise the news as hoaxes at the classification stage. The score threshold can be adjusted to improve sensitivity and specificity in hoax detection. This approach improves accuracy by utilising both methods, reduces the computational burden of the LSTM by pre-screening the news, and improves the efficiency and scalability of the hoax detection system.

5. Conclusion

In this study, a combined approach between lexicon-based and LSTM proved to be very effective in improving the accuracy of hoax news detection on Indonesian news sites with high accuracy results reaching 99%. The use of lexicon to analyse key words and LSTM model to learn complex patterns in news texts has provided satisfactory results in classifying news as hoax or non-hoax. This indicates that a holistic approach to text analysis can produce accurate and relevant results. However, this study also identified some weaknesses that need to be considered in developing hoax detection models. One of the main weaknesses is the reliance on a lexicon that may not include all words relevant to hoax news. In addition, the LSTM model can also suffer from overfitting if not well-tuned, which can reduce performance on test data that has never been seen before.

Acknowledgments

The authors thank to Hasyim Asy'ari University Tebuireng Jombang.

Bibliography

Ade Gohan, M., Andayan, M., Naufal, M., Masliana, M., 2021. Penyuluhan Penyebaran Covid-19

Dengan Pendekatan Participatory Action Research Dalam Menanggapi Berita Hoax Pada Media Sosial. J. IPTEK Bagi Masy. J-IbM 1, 66–73. https://doi.org/10.55537/jibm.v1i2.10

Al-Shabi, M., 2020. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining.

Baidawi, I., 2021. Peranan Pemerintah Kabupaten Situbondo Dalam Menanggulangi Informasi Hoax. Nusant. J. Islam. Stud. https://doi.org/10.54471/njis.2021.2.1.18-24

Balshetwar, S.V., Rs, A., R, D.J., 2023. Fake news detection in social media based on sentiment analysis using classifier techniques. Multimed. Tools Appl. 82, 35781–35811. https://doi.org/10.1007/s11042-023-14883-3

Hutama, L.B., Suhartono, D., 2022. Indonesian Hoax News Classification With Multilingual Transformer Model and BERTopic. Informatica. https://doi.org/10.31449/inf.v46i8.4336

Khanifah, A., Fauzi, A.M., 2022. DAMPAK BERITA HOAX TENTANG COVID-19 TERHADAP PELAKSANAAN PROTOKOL KESEHATAN OLEH MASYARAKAT: (Studi Kasus Group Whatsapp Keluarga). J. Ilm. Din. Sos. 6, 250–267. https://doi.org/10.38043/jids.v6i2.3485

Kominfo, P., n.d. Siaran Pers No.150/HM/KOMINFO/07/2023 tentang Juni 2023, Kominfo Identifikasi 117 Konten Hoaks.

Kurniawan, A.A., Mustikasari, M., 2021. Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia. J. Inform. Univ. Pamulang 5, 544. https://doi.org/10.32493/informatika.v5i4.6760

Ni Made Ayu Juli Astari, Divayana, D.G.H., Indrawan, G., 2020. Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier. J. Sist. Dan Inform. Jsi. https://doi.org/10.30864/jsi.v15i1.332

P, U., Naik, A., Gurav, S., Kumar, A., S R, C., B S, M., 2023. Fake News Detection Using Neural Network, in: 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS). Presented at the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), IEEE, Raichur, India, pp. 01–05. https://doi.org/10.1109/ICICACS57338.2023.10100208

Pardede, J., Ibrahim, R.G., 2020. Implementasi Long Short-Term Memory untuk Identifikasi Berita Hoax Berbahasa Inggris pada Media Sosial. J. Comput. Sci. Inform. Eng. J-Cosine 4, 179–187. https://doi.org/10.29303/jcosine.v4i2.361

Prasetya, F., Ferdiansyah, F., 2022. Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes. J. Sist. Komput. Dan Inform. JSON 4, 132. https://doi.org/10.30865/json.v4i1.4852

Pulungan, H.R., 2022. Perspektif Masyarakat Muslim Angkola Dalam Menyikapi Hoax Melalui Pembelajaran Proses Mental Kognitif. FORUM Paedagog. 13, 1–23. https://doi.org/10.24952/paedagogik.v13i1.4982

Saraswati, R., Nugroho, A.W., Pasaribu, R., 2022. Anti Hoax Movement For Students: Skills Training, Whole Person Education And Technology In Semarang City. SISFORMA 9, 9–17. https://doi.org/10.24167/sisforma.v9i1.3106

Sudrajat, A., Wulandari, R.R., Syafwan, E., 2022. Indonesian Language Hoax News Classification Basedn on Naïve Bayes. J. Appl. Intell. Syst. 7, 70–79. https://doi.org/10.33633/jais.v7i1.5985

Tama, F.R., Sibaroni, Y., 2023. Fake News (Hoaxes) Detection on Twitter Social Media Content through Convolutional Neural Network (CNN) Method. JINAV J. Inf. Vis. 4, 70–78. https://doi.org/10.35877/454RI.jinav1525

Yunanto, R., Purfini, A.P., Prabuwisesa, A., 2021a. Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning. J. Manaj. Inform. Jamika. https://doi.org/10.34010/jamika.v11i2.5362

Yunanto, R., Purfini, A.P., Prabuwisesa, A., 2021b. Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning. J. Manaj. Inform. JAMIKA 11, 118–130. https://doi.org/10.34010/jamika.v11i2.5362

Zhang, J., Dong, B., Yu, P.S., 2020. FakeDetector: Effective Fake News Detection With Deep Diffusive Neural Network. https://doi.org/10.1109/icde48307.2020.00180