RESEARCH ARTICLE

Combating Misinformation: Leveraging Deep Learning for Hoax Detection in Indonesian Political Social Media

Yuliant Sibaroni[1,*], Shuhaimi bin Mahadzir[2], Sri Suryani Prasetiyowati[3], and Aditya Firman Ihsan[4]

[1,3,4]School of Computing, Telkom University, Bandung 40257, Indonesia
[2]Faculty of Engineering, Universiti Teknologi Petronas, Seri Iskandar 32610, Malaysia

*Corresponding email: yuliant@telkomuniversity.ac.id

Abstract: The rampant spread of hoax news in social media, especially in the political domain, poses a significant challenge that requires immediate attention. To address this issue, automatic hoax news detection using machine learning-based artificial intelligence has emerged as a promising approach. With the approaching presidential election in Indonesia in 2024, the need for effective detection methods becomes even more pressing. This research proposed an efficient deep-learning model for detecting political hoax news on Indonesian social media. Word2vec feature representation and three deep learning models – LSTM, CNN, and Hybrid CNN-LSTM – are evaluated to determine the most effective approach. Experimental results revealed that the CNN-LSTM hybrid model outperforms the others, achieving an accuracy of 96% in detecting hoax news on Indonesian social media in the political domain. By leveraging state-of-the-art deep learning techniques, particularly the CNN-LSTM hybrid model, this study contributed to advancing hoax news detection in Indonesia's political landscape. The findings underscore the importance of utilizing sophisticated machine-learning methods to combat the spread of misinformation, particularly during crucial political events such as elections.

Keywords: hoax, deep learning, LSTM, CNN, hybrid, CNN-LSTM, political, word2Vec

1 Introduction

The trend of employing social media for political purposes has witnessed rapid growth in recent years. Social media has become a potent platform for politicians and candidates,

enabling direct communication with their constituents and support base. As the population of social media users continues to expand, politicians can now engage with a broader and more diverse audience than ever before. They leverage features such as live uploads, videos, and stories to disseminate their political messages in real time, stimulate discourse, and mobilize support.

In the forthcoming years, it is projected that the utilization of social media for political endeavors will persist in its ascent, mirroring advancements in technology and shifts in voter behavior. Politicians and political parties must continually adapt to this evolving landscape, recognizing the paramount significance of effective digital communication. Simultaneously, voters are anticipated to become increasingly discerning in evaluating the information they encounter via social media. This heightened scrutiny is essential for making informed political decisions amid the proliferation of fake news and unverified information. Consequently, new challenges arise in ensuring the accuracy and veracity of political information propagated through social media channels.

Researchers have shown keen interest in addressing these challenges, focusing on identifying hoax news across various social media platforms. Researchers have employed diverse sources of social media datasets in their investigations. Several notable examples of research in this domain have explored the issue of fake news on popular platforms such as Facebook [1–5] , Twitter [2, 6–9], Instagram [2] and web-news based [10–14].

Hoax detection can be categorized as one of the research applications in text classification. One method that is superior in text classification is deep learning. One of the strengths of deep learning is that it has automatic feature representation, such as Word2Vec [15] and FastText [16], which enables the formation of highly abstract and complex feature representations from text.

Deep learning can recognize complex patterns and relationships in text. [17, 18]. This includes an understanding of the meaning of words in context, the relationship between words, phrases, and sentences, or shades in language that are often difficult to pick up by other methods [19]. Deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), also enable a good understanding of the hierarchy of information in text.

There is quite a lot of research in identifying hoax news worldwide. Exploration results from papers in several international reputable databases such as Scopus, Science-Direct, IEEE, and ACM Digital Library obtained data that thousands of papers discuss the topic of hoax or fake news from social media data. In contrast, research on identifying hoax news in Indonesia is quite rare. The use of certain aspects of language, such as syntax, politeness [20–22], will make hoaxes in news between one language and another have different characteristics.

The majority of approaches used in identifying hoax news in Indonesian are based on traditional machine learning such as support vector machines (SVM), naïve Bayes, apriori-random forest, K-NN (nearest neighbors), and others [23–25]. The use of a combined apriori and random forest algorithm in [23], was claimed to facilitate researchers in analyzing and evaluating the results of the hoax information detection system obtained. The accuracy results are claimed to be very high, but if we look further, these results are not supported by adequate experiments and good evaluation processes. In hoax news detection, the use of the random forest method outperforms several other classifiers such as gradient boosting, SVM, SGD, AdaBoost, and decision tree [24]. Although random forest performs better than other methods, the results of this research cannot be used as a firm reference because the

size of the dataset used in the research is tiny, below 500 records. Using K-NN combined with term document matrix (TDM) in detecting hoax news by [25] also claimed to provide relatively good results. This model has the advantage of a low computation rate. Similar to previous research, this study also uses a small amount of data.

For deep learning-based approaches with Indonesian datasets, only a few studies have been recorded where the method used is long short-term memory (LSTM) [15] and multi-layer perceptron (MLP) [26]. Regarding the hoax topics studied, the majority of the topics are related to COVID-19, while the other topics are general. Using the LSTM method of Indonesian news hoax detection by [15], shows that the accuracy obtained is relatively good, almost reaching 90% with sufficient data. However, the results' accuracy still has the potential to be improved again to more than 90%. Many studies related to the development of LSTM methods in text classification research have the potential to provide better results in hoax detection, such as gated recurrent unit (GRU), Bi-LSTM, and hybrid LSTM.

Given the absence of prior research addressing this specific political context, our study assesses two deep learning models renowned for their superior performance in text classification: LSTM and CNN. The LSTM method has the advantage of understanding data with long-term dependencies on sequential data, which is a weakness of CNN. CNN has the advantage of detecting local patterns and can perform parallel processing on high-dimensional data. Compared to CNN, LSTM is computationally heavier and more prone to overfitting. Combining these two methods is expected to obtain a combined classifier model superior to the individual models.

A hybrid model, LSTM-CNN, is also tested to determine the most effective deep learning model for identifying political hoax news. Several preprocessing steps, such as stop-word removal and stemming, have been executed, necessitating the utilization of dedicated tools for the Indonesian language. It is noteworthy that the focus on the Indonesian language's political sphere in the context of hoax identification represents a relatively novel area of inquiry in Indonesia.

This research contributes in two significant ways: First, it yields a more robust deep learning model for detecting political hoax news within Indonesian-language social media. Second, it generates a corpus of hoax news specific to the political domain in the context of Indonesian-language social media, further enriching the understanding of this pressing issue.

2   Research Method

2.1   Dataset Collection

The dataset used in this research was collected from the social media Twitter (currently renamed X [27]) with the dataset collection period from December 2022 - March 2023. The keywords used were the names of potential 2024 presidential candidates in Indonesia: Prabowo, Ganjar, and Anies. These keywords were chosen because, before the presidential election in Indonesia in 2024, there would usually be a lot of hoax news aimed at attacking presidential candidates from the opposing camp. The initial number of datasets was around 12 thousand news data, and the data taken was from the prominent tweets that contained truth values. After data preprocessing and balancing, the number of data labeled hoax and non-hoax is 5,395 each. Data was labeled as a hoax if it contained information that was not true or not based on facts; otherwise, it would be categorized as non-hoax. Data

labeling was performed by three people, and when different labeling occurs, the labeling process is carried out using majority voting. The labelers were bachelor's degree graduates in computer science. Examples of hoax-labeled data can be seen in Table 1. As can be seen in the data examples in Table 1, the use of language in Twitter social media (X) looks informal, contains many abbreviations, and the content is partly in the form of sarcasm or insults

Table 1: Example hoax data on Twitter (X)

Example Hoax Tweet (X)
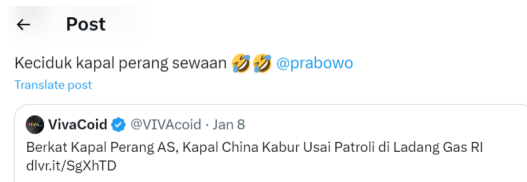
caught on a chartered warship
This post states that the patrol boat is a chartered boat. In fact, there is no official news stating this.

yes, it's wrong that DKI rice is lumped in which other provinces are shared
This relates to the case of rice distribution in DKI Jakarta during the Covid-19 pandemic, some of which was not distributed because it was damaged. In fact, this rice was not intentionally stored by the government.
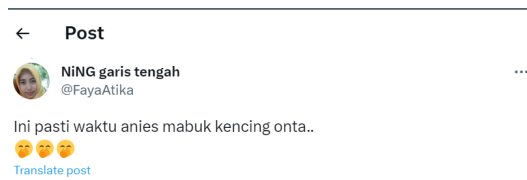
This must be when anies was drunk on camel piss
This sentence is a type of insulting sentence and in fact: this is not happening.

2.2   Preprocessing and Labelling

The raw dataset obtained then entered the preprocessing stage, which included case folding, tokenization, stopword removal, stemming, cleaning from duplicates, and cleaning

against null data rows. Case folding was changing the font type to be in the form of all lowercase letters. This was done so that features with the same meaning but presented in different fonts were still considered the same feature. In addition to strengthening a particular word feature, this process could reduce the number of word features extracted. Tokenization was conducted to break the sentence into word tokens. Next, the Stopword removal process was carried out, which was a process to eliminate words that were considered unimportant. The stopword dictionary used in this research was based on the stopword dictionary in the Natural Language Toolkit (NLTK) library developed by a community of researchers in the field of natural language processing (NLP). The next step was stemming, which removed prefixes and suffixes from a word into its basic form (root word). The goal was to produce a more general form of words in the text so that words that vary could be considered the same. The stemming process was performed for the Indonesian dataset used in this research using the Sastrawi library [28] a simple Python library that converted Indonesian words to their base form.

2.3   Deep Learning Model

This research utilized three deep learning architectures: LSTM, CNN, and the CNN-LSTM Hybrid model. Prior to delving into the detailed discussion of the social media hoax news classification process, let us first explore each of these deep learning models.

2.4   LSTM Model

The first architecture was LSTM, which was used by [15] in news hoax detection on websites. The method in the study was the baseline of this research. LSTM is a type of artificial neural network architecture that belongs to the recurrent neural networks (RNN) category. LSTM was developed to address the problem of long-term dependencies in sequence data and is particularly effective in natural language processing, speech recognition, and time series modeling tasks.

LSTMs had memory units called "memory cells" that keep information from previous inputs over multiple time steps. This allowed them to overcome the memory problems commonly encountered in conventional RNNs. LSTMs had three main gates [29]:

1. Forget Gate: This gate decides which information should be removed from the memory cell. This helps in coping with distance length changes in the data.
2. Input Gate: This gate decided which information should be added to the memory cell based on the current input.
3. Output Gate: This gate generated the output based on the memory cell that has been updated and the current input.

The combination of these gates allowed the LSTM to manage and process sequences of data well, which made it very useful in NLP tasks such as text classification, sentiment analysis, and text generation.

2.5   CNN Model

Another deep learning architecture under consideration was CNN, which drew inspiration from the human brain's visual information processing mechanisms. CNNs feature

convolutional layers designed to extract local features from input data. These layered incorporate convolution kernels or filters to identify patterns and features within the data. Pooling layers were also employed to reduce data dimensionality and eliminate less significant features, enabling CNNs to automatically extract essential features from data, making them particularly adept in image recognition tasks. In the context of Text Classification, CNNs were adapted with slight modifications. After converting data into text vectors using Word2Vec, a padding process was performed to standardize text length by adding padding tokens (0 tokens) to achieve uniformity [30]. A typical CNN model had the following components [31]:

1. Convolutional Layer
   The convolutional layer is the core component in CNN that serves to extract features from the input data.
2. Pooling Layer
   This layer is used to reduce the data dimension and eliminate less important features.
3. Hidden Layers
   The results of the previous 2 layers then go into additional hidden layers, such as the Dense layer commonly used in neural networks.
4. Output Layer
   The output layer is the last layer in the CNN model that produces a prediction or output.

2.6   CNN-LSTM Hybrid Model

This study's third deep learning architecture combined the strengths of CNN's spatial feature recognition with LSTM's sequence and context understanding capabilities. This hybrid model was constructed in the following manner.

1. Input Layer
   The input layer received the sequential input data, which, in the case of text classification, consisted of word embeddings or word vectors representing the words in the text.
2. Initial Feature Extraction with CNN
   CNN was used as the first layer to retrieve local features from the text. This meant the text would be divided into small parts, and features such as word patterns or subwords would be extracted from each. The feature map generated by the convolution and pooling layers of the CNN would be the initial representation of the text.
3. Max Pooling Layer
   The max pooling layer downsampled the feature maps obtained from the convolutional layers by selecting the maximum value within each region of the feature map. This helped in reducing the dimensionality of the feature maps while retaining the most critical information.
4. LSTM Layer for Sequence Understanding
   The resulting feature map from the CNN was loaded into the LSTM layer. The LSTM helped to understand the sequence of words in the text and process them in context. LSTMs had memory cells that allowed them to maintain information from previous inputs in the text sequence, which was helpful in overcoming the problem of remote dependency in the text sequence.

5. Fully Connected Layers
   The output of the LSTM layer was flattened or pooled and then fed into one or more fully connected layers. These fully connected layers perform classification based on the learned features from the input text. They mappeda the features to the output classes using activation functions such as softmax for multi-class classification.
6. Output Layer
   The output layer produced the final classification output, which could be the predicted class label or a probability distribution over the classes.

2.7 Hoax News Identification using Deep Learning

The hoax news identification process stages were carried out by classifying hoax news using deep learning. The stages were as follows:

1. Text Vectorization
   The text dataset was converted into a vector representation using Word2Vec (in this research, one word was explored by embedding technique, word2vec) so that it could be used as input for the LSTM model.
2. Data Splitting
   The dataset was divided into 2, namely training data and testing data with the composition of train data: test data was 90: 10.
3. CNN Modeling
   - Convolutional Layer: extracted features from input data using Word2Vec.
   - Pooling Layer: Reduced or eliminated less important features using Max Pooling.
   - Hidden Layers: using fully connected architecture
   - Output Layers using softmax activation function
4. LSTM Modeling:
   The LSTM model built has the following layers structure:
   - Input Layer: text sequence vector based on text dataset using word2vec.
   - Hidden layers using fully connected layers.
   - Output Layers using softmax activation function
5. CNN-LSTM Hybrid Modeling
   - Process the dataset until a feature map was formed using Convolutional and pooling layers in the CNN architecture.
   - Processing the feature map resulting from CNN modeling to the hidden and output layers of LSTM
   - CNN-LSTM hybrid model architecture, can be seen in Figure 1.
6. Model Training and Hyperparameter Tuning The optimal learning rate for training the CNN-LSTM model in this research was determined by hyperparameter tuning, where different learning rates were tried and evaluated on the validation set. The hyperparameter tuning of CNN-LSTM models was optimized using the Adam algorithm. Adam optimization was a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments.
7. Model Evaluation
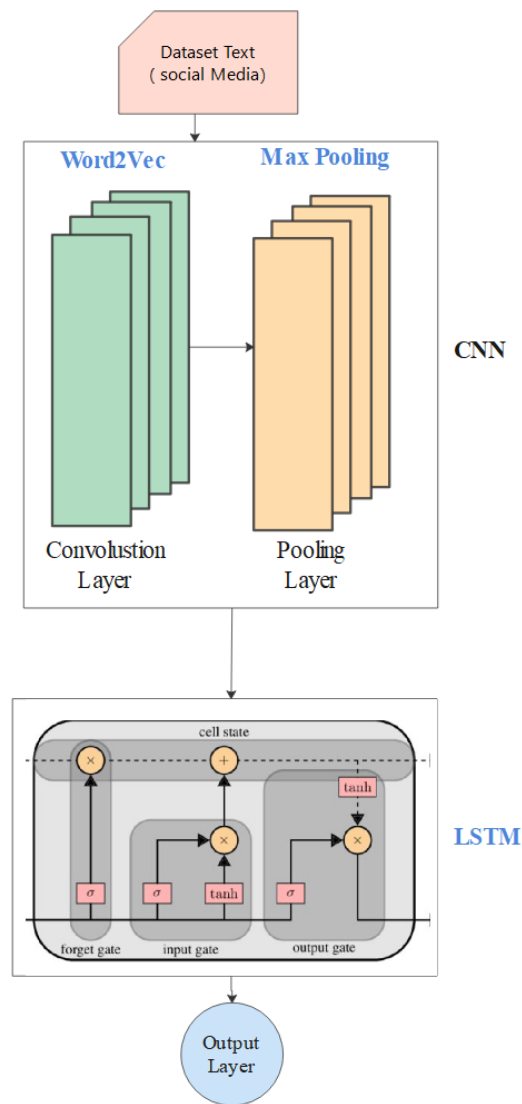   The LSTM model evaluation process used a 10-cross validation mechanism where

Figure 1: Architecture of hybrid deep learning model (CNN-LSTM) for hoax news detection on political Indonesian social media.

the data would be split into ten equal-sized subsets called folds. The dataset was randomly divided into ten disjoint subsets of approximately equal size. Each subset contained an equal proportion of samples from the original dataset. The cross-validation process was repeated ten times, each iteration using a different fold as the test set and the remaining nine as the training set. For each iteration, the model was trained on the training set (9 folds) and evaluated on the test set (1 fold).

3   Result

We conducted training sessions for three deep learning models: LSTM, CNN, and a Hybrid CNN-LSTM model, each for ten epochs. We evaluated their performance and present the results in Table 2. In this experiment, it can be shown that on ten fold data, the majority of the CNN-LSTM hybrid model is superior to other deep learning models (based on accuracy in test data).

Table 2: Hoax classification accuracy for various epoch values

| epoch | Model Accuracy | | | | | |
| | LSTM | | CNN | | Hybrid | |
	train	test	train	test	train	test
1	0.703	0.735	0.611	0.784	0.610	**0.810**
2	0.800	0.817	0.765	0.888	0.770	**0.904**
3	0.857	0.848	0.857	0.924	0.869	**0.941**
4	0.898	0.851	0.882	0.935	0.894	**0.938**
5	0.914	0.850	0.912	**0.954**	0.904	0.950
6	0.932	0.898	0.919	0.950	0.929	0.958
7	0.937	0.900	0.923	**0.956**	0.930	0.953
8	0.953	0.909	0.936	**0.955**	0.935	0.954
9	0.952	0.905	0.942	0.956	0.934	**0.962**
10	0.956	0.882	0.937	0.965	0.944	**0.968**

The accuracy assessment of hoax news classification in the political domain using the LSTM model is depicted in Figure 2. Initially, during the first epoch, both the training and validation accuracies were relatively low, hovering around 70%. Subsequently, the accuracy of the training steadily increased with each epoch until the 10th. However, the validation accuracy plateaued at the eighth epoch, reaching a maximum of 90.8%. This observation implies that the optimal epoch count for the LSTM model is 8, achieving a test data accuracy of 90.8%. The accuracy patterns for training and validation data are graphically illustrated in Figure 2, further highlighting the optimal performance at the eighth epoch, where the validation accuracy reached approximately 90.9%.

Turning to the CNN model in Figure 3, training and test accuracies were lower in the initial epoch than LSTM, at approximately 60%. However, the training accuracy consistently increased and peaked at the ninth epoch, coinciding with a test data accuracy of 95.6%. Notably, the CNN model demonstrated a more stable accuracy pattern, with training accuracy closely matching test accuracy starting from the fifth epoch.

Hybrid CNN-LSTM model performance was quite similar to the CNN model; in the first epoch, the training and test accuracies were around 60%, lower than the LSTM model. Nevertheless, the training accuracy exhibited steady growth, culminating at the 10th epoch with a remarkable test data accuracy of 96.7%. This value represented the highest accuracy among the three models. The accuracy pattern of the Hybrid CNN-LSTM model in Figure 4 is characterized by its consistency, as both training and validation accuracies steadily increased, peaking at the 10th epoch.
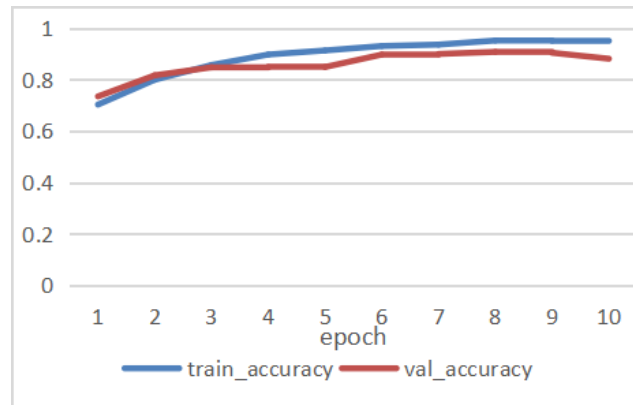
Figure 2: Accuracy pattern of LSTM model for various epoch values.
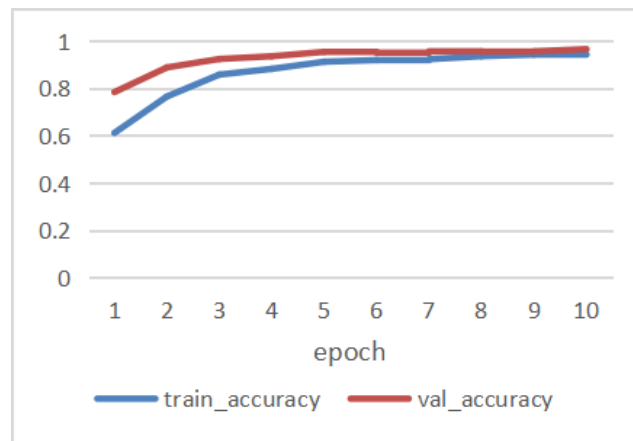

Figure 3: Accuracy pattern of CNN Model for various epoch values.

4 Discussion

Compared to the other two models, the Hybrid CNN-LSTM model achieved the highest training data accuracy with an optimal epoch count of 10, resulting in an accuracy of 96.7%. Based on the achieved accuracy values, it can be inferred that the CNN-LSTM hybrid model was the most optimal deep learning model for detecting political hoax news on social media. Notably, this model exhibits significantly superior accuracy performance compared to the LSTM model, which had previously been employed in hoax news detection research in Indonesia. However, the performance gain was relatively small if the CNN-LSTM hybrid model was compared to the CNN model alone. Even so, it was still concluded that the CNN-LSTM hybrid model was still the best deep learning model in this case. From these results, it could be concluded that the output of the convolution layers of the CNN model would provide more optimal results if processed using the LSTM model than processed on fully connected layers as in the full CNN model. It is important to note that the dataset
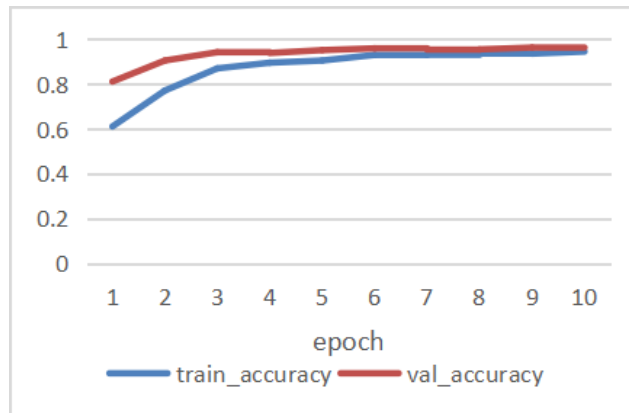
Figure 4: Accuracy pattern of hybrid CNN-LSTM for various epoch values.

used for this study was sourced from social media, and the news content was specific to the political domain, distinguishing it from prior research that employed more general datasets and non-social media sources.

5   Conclusion

This research is centered around detecting political hoax news circulating on Indonesian social media platforms, which has seen limited exploration with Indonesian language datasets thus far. The unique characteristics of news dissemination on websites and social media, coupled with the distinct attributes of political hoaxes, set this research apart from previous investigations.

In this study, we adopted a deep learning-based approach, specifically LSTM, as the baseline method. This LSTM method was initially designed to detect hoax news within the context of website news, addressing a more diverse set of cases. Consequently, we evaluated three deep learning models for identifying political hoax news within Indonesian social media data: LSTM, CNN, and a CNN-LSTM hybrid model. We aimed to identify the most effective and efficient deep learning model for detecting political hoax news within social media.

The experimental results consistently indicate that the CNN-LSTM hybrid model outperforms both the LSTM and CNN models regarding accuracy and stability. The CNN model, which draws inspiration from deep learning techniques initially designed for image data classification, also demonstrates superior performance compared to LSTM, the baseline method utilized in this study. Notably, the CNN-LSTM hybrid model attains the highest accuracy, reaching an impressive 96%, signifying a significant improvement over the LSTM baseline model, which achieves an accuracy of 91%.

Acknowledgments

References

[1] D. Chaudhari, K. Rana, R. Tannu, and S. Yadav, "Fake news detection on social media," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 1132–1137, July 2020.

[2] S. R. Sahoo and B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, Mar. 2021.

[3] A. Souri, S. Hosseinpour, and A. M. Rahmani, "Personality classification based on profiles of social networks' users and the five-factor model of personality," *Human-centric Computing and Information Sciences*, vol. 8, p. 24, Dec. 2018.

[4] A. Galli, E. Masciari, V. Moscato, and G. Sperlí, "A comprehensive benchmark for fake news detection," *Journal of Intelligent Information Systems*, vol. 59, pp. 237–261, Aug. 2022.

[5] S. Preston, A. Anderson, D. J. Robertson, M. P. Shephard, and N. Huhe, "Detecting fake news on Facebook: The role of emotional intelligence," *PLOS ONE*, vol. 16, p. e0246757, Mar. 2021.

[6] G. Bharath, K. J. Manikanta, G. B. Prakash, R. Sumathi, and P. Chinnasamy, "Detecting fake news using machine learning algorithms," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, (Coimbatore, India), pp. 1–5, IEEE, Jan. 2021.

[7] D. F. Brianna, E. Surya Negara, and Y. N. Kunang, "Network centralization analysis approach in the spread of hoax news on social media," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, (Batam Island, Indonesia), pp. 303–308, IEEE, Oct. 2019.

[8] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid deep learning models for sentiment analysis," *Complexity*, vol. 2021, pp. 1–16, Aug. 2021.

[9] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid CNN and RNN models," in *Proceedings of the 9th International Conference on Social Media and Society*, (Copenhagen Denmark), pp. 226–230, ACM, July 2018.

[10] P. Akhtar, A. M. Ghouri, H. U. R. Khan, M. Amin Ul Haq, U. Awan, N. Zahoor, Z. Khan, and A. Ashraf, "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions," *Annals of Operations Research*, vol. 327, pp. 633–657, Aug. 2023.

[11] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive Bayes: Aggregating one-dependence estimators," *Machine Learning*, vol. 58, pp. 5–24, Jan. 2005.

[12] A. Glazkova, M. Glazkov, and T. Trifonov, "g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation* (T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, eds.), vol. 1402, pp. 116–127, Cham: Springer International Publishing, 2021.

[13] S. Al-Maadeed, A. Hassaine, A. Bouridane, and M. A. Tahir, "Novel geometric features for off-line writer identification," *Pattern Analysis and Applications*, vol. 19, pp. 699–708, Aug. 2016.

[14] M. I. Nadeem, K. Ahmed, Z. Zheng, D. Li, M. Assam, Y. Y. Ghadi, F. H. Alghamedy, and E. T. Eldin, "SSM: Stylometric and semantic similarity oriented multimodal fake news detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, p. 101559, May 2023.

[15] R. Yusuf and S. Suyanto, "Hoax detection on Indonesian text using long short-term memory," in *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, (Yogyakarta, Indonesia), pp. 268–271, IEEE, Aug. 2022.

[16] A. Gupta, R. Sukumaran, K. John, and S. Teki, "Hostility detection and covid-19 fake news detection in social media," Jan. 2021. arXiv:2101.05953 [cs].

[17] H. Byun and S.-W. Lee, "A survey on pattern recognition applications of support vector machines," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, pp. 459–486, May 2003.

[18] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis," *Informatics*, vol. 8, p. 79, Nov. 2021.

[19] S. Saha, Z. Gan, L. Cheng, J. Gao, O. L. Kafka, X. Xie, H. Li, M. Tajdari, H. A. Kim, and W. K. Liu, "Hierarchical deep learning neural network (HiDeNN): An artificial intelligence (AI) framework for computational science and engineering," *Computer Methods in Applied Mechanics and Engineering*, vol. 373, p. 113452, Jan. 2021.

[20] A. Izadi and A. Jalilifar, "Introduction: Politeness and impoliteness research," *Journal of Research in Applied Linguistics*, vol. 13, Sept. 2022.

[21] A. L. Marciotto Oliveira and M. Miranda, ""Calling a spade a spade": Impoliteness and shame on twitter," *Journal of Research in Applied Linguistics*, vol. 13, Sept. 2022.

[22] H. Ghanbar and R. Rezvani, "Research questions in applied linguistics research: A microscopic analysis of their distributional and syntactical aspects," *Journal of Research in Applied Linguistics*, vol. 14, Apr. 2023.

[23] M. P. Utami, O. D. Nurhayati, and B. Warsito, "Hoax information detection system using apriori algorithm and random forest algorithm in twitter," in *2020 6th International Conference on Interactive Digital Media (ICIDM)*, (Bandung, Indonesia), pp. 1–5, IEEE, Dec. 2020.

[24] F. W. Wibowo, A. Dahlan, and Wihayati, "Detection of fake news and hoaxes on information from web scraping using classifier methods," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, (Yogyakarta, Indonesia), pp. 178–183, IEEE, Dec. 2021.

[25] E. Zuliarso, M. T. Anwar, K. Hadiono, and I. Chasanah, "Detecting hoaxes in Indonesian news using TF/TDM and k nearest neighbor," *IOP Conference Series: Materials Science and Engineering*, vol. 835, p. 012036, Apr. 2020.

[26] A. Rusli, J. C. Young, and N. M. S. Iswari, "Identifying fake news in Indonesian via supervised binary text classification," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, (Bali, Indonesia), pp. 86–90, IEEE, July 2020.

[27] M. Binder, "Twitter's rebrand to X has its website looking like a mess," July 2023.

[28] "Sastrawi: Library for stemming Indonesian (Bahasa) text."

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[30] K. Radhika, K. R. Bindu, and P. Latha, "A text classification model using convolution neural network and recurrent neural network," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 15, pp. 1549–1554, 2018.

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov. 1998.