



## Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter

Muhammad Ikram Kaer Sinapoy<sup>1</sup>, Yuliant Sibaroni<sup>2</sup>, Sri Suryani Prasetyowati<sup>3</sup>

<sup>1,2,3</sup>School of Computing, Informatics, Telkom University

<sup>1</sup>ikramsinapoy@student.telkomuniversity.ac.id, <sup>2</sup>yuliant@telkomuniversity.ac.id, <sup>3</sup>srisuryani@telkomuniversity.ac.id

### Abstract

*In recent years, social media users have been increasing significantly, in January 2022 social media users in Indonesia reached 191 million people which has an increase of 12.35% from the previous year as many as 170 million people, With this massive increase every year, more and more people tend to seek and consume information through social media. Despite the many advantages provided by social media, However, the quality of information on social media is lower than in traditional news media there is a lot of hoax information spreading. With many disadvantages felt by hoax information, it has led to many research to detect hoax information on social media, especially information that is widely spread on Twitter. There are several previous researches that use various models using machine learning and also using deep learning to detect hoax. deep learning is very well used to perform several text classification tasks, especially in detecting hoax. The aim of this paper is to compare the LSTM and IndoBERT methods in detecting hoax using datasets taken from Twitter. In this study, two experiments work are conducted, LSTM and IndoBERT methods. The experimental results is average value obtained from experiments using 10-fold cross-validation. The IndoBERT model shows good performance with an average accuracy value of 92.07%, and the LSTM model provides an average accuracy value of 87.54%. The IndoBERT model can show good performance in hoax detection tasks and is shown to outperform the LSTM model which can provide the best average accuracy results in this study.*

*Keywords: hoax detection; social media; twitter; deep learning; LSTM; IndoBERT*

### 1. Introduction

In recent years, social media users have been increasing significantly, in January 2022 social media users in Indonesia reached 191 million people which has an increase of 12.35% from the previous year [1], With this massive increase every year, more and more people tend to seek and consume information through social media [2], The reason people get information through social media is because quite easy and faster to get information compared with getting information through traditional media such as newspapers. Despite the many advantages provided by social media, However, the quality of information on social media is lower than in traditional news media there is a lot of hoax information spreading, this is due to the easy access and also the lack of control of the internet [3]. According to data from the Ministry of Communication and Information of the Republic of Indonesia (KEMKOMINFO), the spread of hoax on social media in the last three years has occurred in as many as 9,546 cases. Hoax information is intentionally written to mislead readers [4] and it carried out by irresponsible people with various purposes, such as the purpose of a group or individual that can have a

serious negative impact on society [2]. With many disadvantages felt by hoax information, it has led to many research to detect hoax information on social media, especially information that is widely spread on Twitter [5].

There are several previous researches that use various models using machine learning [6]–[8] and also using deep learning [9]–[11] to detect hoax. deep learning is very well used to perform several text classification tasks, especially in detecting hoax [12]. One of the commonly used models is RNN (Recurrent Neural Network), RNN developed to overcome subsequent data, but RNN has limitations in capturing long-term dependencies. Long Short-Term Memory (LSTM) is a model of RNN that developed or modified to overcome the limitations of RNN in capturing long-term dependencies by remembering long-term information [13], in other words, the limitations of long-term dependencies in RNN, are not a problem in LSTM, so LSTM is more efficient in process, prediction, and classification of data. In research [13] to detect hoax in Indonesian languages using the Long Short-Term Memory (LSTM) method, the results of the research

obtained the average value of precision, recall, f1 score 0.819; 0.809; and 0.807 from several experiments. In another research [9] to analyze hoax on Indonesian news using several deep learning models, the LSTM model produces a high accuracy value of 95.6% without using dropout, after using dropout the accuracy value has increased slightly to 95.6%.

Recent studies have shown that pre-trained models trained on a large corpus can successfully accomplish various tasks with transfer learning. One of the pre-trained models that can provide the best results for specific tasks is the BERT architecture [14]. IndoBERT is a transformer-based model using BERT [14]. There are currently two kinds of IndoBERT, developed by IndoNLU and IndoLEM. IndoBERT from IndoNLU conducts the training process with datasets collected from various sources such as social media, blogs, news, and websites [15]. IndoBERT from IndoLEM conducts the training process with datasets from Wikipedia Indonesia, Indonesian news articles, and the Indonesian Corpus website [16]. In research [14] using the IndoBERT model trained on the Indonesian language corpus, IndoBERT fine-tuned models, IndoNLU and IndoLEM provide the same performance with accuracy 97.67. IndoBERT outperformed the other fine-tuned models and the multilingual BERT model with the uncased version.

LSTM and IndoBERT are methods designed to process sequential input data, such as natural language, and both methods have often been used in text classification tasks, especially in hoax detection, and the LSTM and IndoBERT methods have shown good accuracy in hoax detection tasks. because these two methods have shown good performance results in hoax detection tasks, this study will compare the accuracy results of the LSTM and IndoBERT methods in detecting hoax using datasets taken from Twitter.

## 2. Research Methods

### 2.1 System Flowchart

The design of the work system in this study as shown in Figure 1, begins with crawling the dataset taken from the tweets of Twitter users, and continued with the manual labeling process to determine whether a tweet is a hoax or not, then it will correct the structure of inconsistent sentences by doing pre-processing. After pre-processing the dataset, then the next process in the LSTM model is carried out the training process using Word2Vec before splitting the data for training data and test data using K-fold cross-validation, in the IndoBERT model does not need to do the Word2Vec process and directly splits the data using K-fold cross-validation, then the training and testing process will be carried out. After all the processes are completed, the last step in this research will compare the accuracy

results obtained from the LSTM and IndoBERT models in detecting hoaxes.

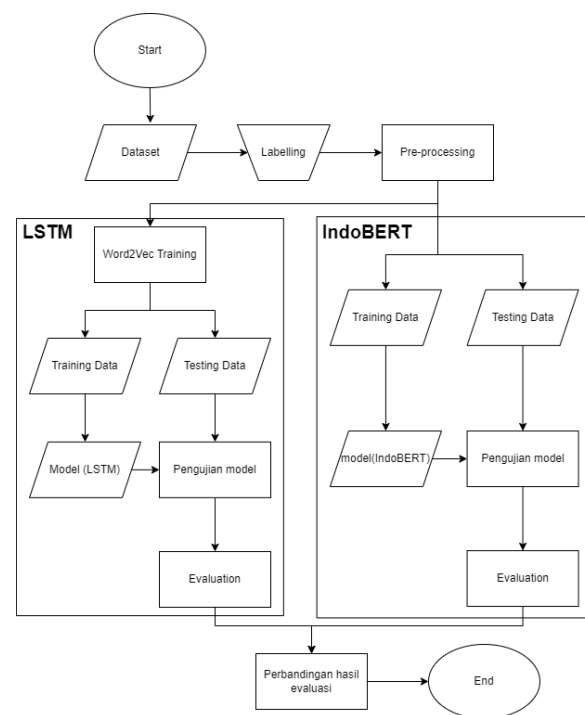


Figure 1. Flowchart

### 2.2 Data

The dataset used in this study was taken from Twitter by crawling using the snsrape library, this library can collect data from Twitter with various features that we need for our research. the data taken is Indonesian tweets with the keywords "covid-19" and "corona", with a time range of data collection from March 2020 - May 2020. Table 2, is an overview of the features of the data taken from Twitter.

Table 1. Description of Features in The Data

Feature	Description
Conversation ID	The ID of a conversation from a tweet post
Date	The date of a tweet post
User ID	The ID of the user who posted the tweet
Username	A name that can be used as a reference for a user account, and each account cannot use the same Username.
Tweet	Message posted to Twitter that can contain photos, videos, links, and text.
Language	The language used in each tweet post.
Location	Location of the tweet upload
Number of replies	The number of replies to a tweet post
Number of retweets	The number of retweets to a tweet post
Number of likes	The number of likes to a tweet post

### 2.3. Labelling

Labeling is a process to determine or identify tweet data including hoax information or not which is done manually by the author. The author labels about 3000

data from all existing data. The results of data comparison with hoax and non-hoax labels from existing datasets are shown in table 2.

Table 2. Labeling Result

Label	Amount	Percentage
Hoax	1.404	47%
Non-Hoax	1.599	53%

## 2.4 Pre-processing

Pre-processing is used to process datasets into information that is more efficient and useful in carrying out the classification process. datasets taken by crawling from twitter usually often have errors, missing values, and inconsistencies. Pre-processing consists of several stages, first, remove emoji serves to clean some emoji or emoticons in the tweet data. After that, case folding aims to convert all letters into lowercase letters, and only letters 'a' to 'z' are accepted other than those that are considered delimiters. then, normalization which aims to correct abbreviated or unclear words that are matched in the normalization dictionary of words. Next is clean tweet, in this section function to eliminate things that are not too important that can affect the results of research such as punctuation, hashtags, and links. After that, the stemming process is to convert the word into its basic form. and the last step is filtering which aims to remove words that are considered not important.

Table 3. Pre-processing Flow

Process	Sentence
Initial	b'@abu_waras @mansurajjah434 Semoga klrge, kawan2 dan org yg mengenali dijauhkan dri wabak Corona ni 🙏🤔'
Remove emoji	b'@abu_waras @mansurajjah434 Semoga klrge, kawan2 dan org yg mengenali dijauhkan dri wabak Corona ni'
Case folding	b'@abu_waras@mansurajjah434 semoga klrge, kawan2 dan org yg mengenali dijauhkan dri wabak corona ni'
Normalization	b'@abu_waras@mansurajjah434 semoga keluarga, kawan dan orang yang mengenali dijauhkan dari wabah corona ini'
Clean Tweets	semoga keluarga kawan dan orang yang mengenali dijauhkan dari wabah corona ini
Stemming	moga keluarga kawan dan orang yang kenal jauh dari wabah corona ini
Filtering	moga keluarga kawan orang kenal jauh wabah corona

## 2.5 K-Fold Cross Validation

The training and evaluation process of this study uses K-fold cross-validation, the use of cross-validation to test the robustness of the model in the classification of unseen data [17]. k value used in k-fold cross-validation is 10. The dataset used in this study will be randomly divided into 10 equal parts (folds) which are repeated ten times. The dataset, 90% is used for training and 10% is used for testing. By shifting the folds at each iteration,

new training data and testing data will be generated. Figure 2 shows the distribution of data in training and testing using 10-fold cross-validation.

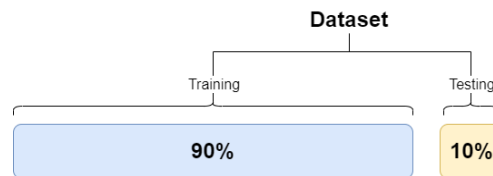


Figure 2. 10-Fold Cross Validation

## 2.6 Word2Vec

Word2Vec is the name of the word vector representation, two architectures in Word2Vec modeling are used to represent word vectors, which are continuous bag-of-words (CBOW) and skip-gram. In this research, we will use a skip-gram architecture that uses the current word as a target for the neighbor words. In building the Word2Vec model, Figure 3 shows three processes are involved, vocabulary builder, context builder, and neural network (skip-gram architecture) [11].

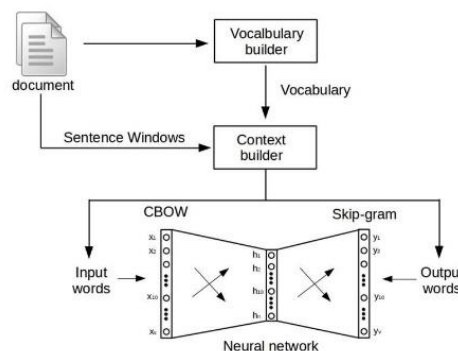


Figure 3. Word2Vec Architecture

The first part of the Word2Vec model is the vocabulary builder, which is used to build the vocabulary of the corpus text. This section will collect all the unique words from the corpus and use to build a dictionary. The vocabulary builder process will result in a dictionary of words with word indexes and occurrence values for each word [11].

Context builder is a process to find the relation between the occurrence of one word and other words around it by using the context window concept or commonly called a sliding window [11]. The Word2Vec training process in this study uses one type of architecture from Word2Vec, that is skip-gram which will predict the context or word (output) around the current word (input) which is bounded by a window. The window is used to obtain the input and target words, window will be moved from the beginning to the end of the word order. An illustration of the window can be seen in Figure 4.

Text	Word Pairs
virus corona adalah penyakit menular	(virus, corona) (virus, adalah)
virus <u>corona</u> adalah penyakit menular	(corona, virus) (corona, adalah) (corona, penyakit)
virus corona <u>adalah</u> penyakit menular	(adalah, virus) (adalah, corona) (adalah, penyakit) (adalah, menular)
virus corona adalah <u>penyakit</u> menular	(penyakit, corona) (penyakit, adalah) (penyakit, menular)
virus corona adalah penyakit <u>menular</u>	(menular, penyakit) (menular, adalah)

Figure 4. Illustration of The Window

This process is used to perform training so that each word can be represented with a vector. Word2Vec uses an artificial neural network architecture that uses 3 layers, input layer, hidden layer, output layer [11] formed from skip-gram architecture. Figure 5 shows the skip-gram architecture to generate word2vec.

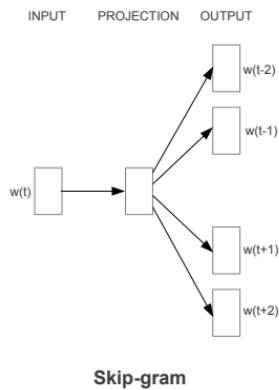


Figure 5. Skip-gram Architecture

## 2.7 LSTM

One of the commonly used models that is a modification of RNN is Long Short-Term Memory (LSTM). LSTM is modified to complement the shortcomings of RNNs that cannot predict words and remember information stored for long periods of time and delete data that is no longer needed [7]. In LSTM, there are some parts that control the use and update of previous information, namely the input gate, forget gate, and output gate and a memory cell [13]. The memory cell and the three gates are designed to read, store, and update past information. Figure 6 shows the LSTM architecture.

There are several steps in the LSTM model in processing input data, the initial stage starts through the *forget gate* ( $f_t$ ). At this stage, parts that are not needed or have less meaning in this case will be removed. The calculation of the *forget gate* value uses the equation (1)

$$f_t = \sigma([h_{(t-1)}, x_t] + b_1) \quad (1)$$

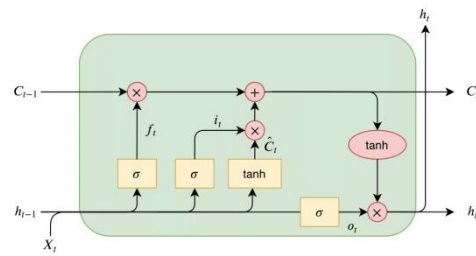


Figure 6. LSTM Architecture

In the next step, the data or information is processed through the *input gate* ( $i_t$ ) using equation (2), this process will divide and determine the information that will be updated to the cell state section, and in this step also creates a new candidate vector which will then be added to the section *cell state* ( $C1_t$ ) using equation (3).

$$i_t = \sigma([h_{(t-1)}, x_t] + b_2) \quad (2)$$

$$C1_t = \tanh(W \cdot [h_{(t-1)}, x_t] + b_3) \quad (3)$$

After that, update the value of the old cell state ( $ct-1$ ) to the new cell state ( $c_t$ ) using equation (4).

$$c_t = f_t * C_{(t-1)} + i_t * C1_t \quad (4)$$

The last step is in the output gate section, after producing sigmoid and tanh output values, the results will be multiplied before going to the next step using equations (5) and (6). After all the calculation processes in the LSTM model are complete, it will produce a classification value.

$$o_t = \sigma(W \cdot [h_{(t-1)}, x_t] + b_4) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

## 2.8 IndoBERT

IndoBERT is the result of a modification of BERT Base that follows the settings of BERT-Base (uncased)[18]. In the process, IndoBERT uses the mechanism of transformers, where the mechanism learns the relationship between each word in a sentence. IndoBERT uses two mechanisms, an encoder to read the input and a decoder to generate predictions [18]. Unlike other language models that can only read input text sequentially from left to right or vice versa, using the BERT method can read the entire word at once [19]. There are two steps used in IndoBERT [20], namely pre training, during the pre-training process the model is trained on unlabeled data on different pre-training tasks. And fine-tuning, the pre-trained parameters are initialized to the BERT model, and the parameters are fine-tuned with labeled data from the given task. Each given task has a separate fine-tuned model, although it is initialized with the same pre-trained parameters. Currently, there are two types of IndoBERT trained on different corpus, one of which is IndoNLU [14]. IndoNLU includes 12 tasks, which perform the trained process using a large Indonesian clean dataset called Indo4B which is collected from



various public sources such as social media, blogs, news, and websites [15].

### 2.9 System Performance Measure

System performance measures in this study using confusion matrix. The confusion matrix contains information about the actual classification as well as the predictions that have been made by the classifier model, which aims to analyze the impact of each scenario on the performance of an analysis model [21]. Table 4 shows the confusion matrix.

Table 4. Confusion Matrix

	Actual class (+)	Actual class (-)
Predicted class (+)	True Positive(TP)	False Positive(FP)
Predicted class (-)	False Negative(FN)	True Negative(TN)

The confusion matrix measurement in this study uses the accuracy value. Accuracy is a measure of how much data is correctly predicted in the classification process. Accuracy uses equation (7).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

## 3. Results and Discussions

This section explains the performance results using the LSTM and IndoBERT methods that have been implemented in building a system to detect hoaxes from datasets taken on Twitter, and also shows a comparison of the average results obtained from the LSTM and IndoBERT methods.

### 3.1 LSTM

The results from applying the LSTM method to the dataset can be seen in table 5.

Table 5. Accuracy of Each Fold Using LSTM

Fold	Accuracy
1	<b>90.69</b> %
2	88.70 %
3	86.37 %
4	87.00 %
5	89.99 %
6	87.00 %
7	85.66 %
8	85.66 %
9	88.33 %
10	86.00 %

In table 5, it can be seen from the use of the LSTM method on the dataset, the highest accuracy value is obtained from fold 1 with a value of 90.69%, the lowest accuracy value in fold 7 and 8 is 85.66%, while the precision value is 84.33%, the recall value is 86.66% and the average accuracy result is 87.54%.

### 3.2 IndoBERT

The results from applying the IndoBERT method to the dataset can be seen in table 6.

Table 6. Accuracy of Each Fold Using IndoBERT

Fold	Accuracy
1	92.02 %
2	96.67 %
3	95.01 %
4	<b>98.66</b> %
5	<b>98.66</b> %
6	97.33 %
7	<b>98.66</b> %
8	96.33 %
9	94.00 %
10	53.33 %

In table 6, it can be seen from the use of the IndoBERT method on the dataset, the highest accuracy value is obtained from fold 7 with a value of 98.66%, the lowest accuracy value in fold 10 is 53.33%, while the precision value is 93.33%, the recall value is 97.22% and the average accuracy result is 92.07%.

### 3.3 Comparison of Evaluation Results

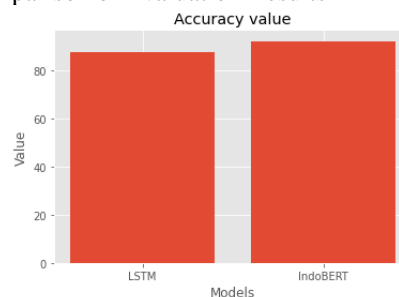


Figure 7. Comparison of Accuracy Result

Figure 7 shows the performance graph of the LSTM and IndoBERT models in this study, the value used in the graph in Figure 7 uses the average value obtained from experiments using 10-fold cross-validation. The IndoBERT model shows good performance with an average accuracy value of 92.07%, and the LSTM model provides an average accuracy value of 87.54%.

Figure 8 shows the results of the average loss value of the LSTM and IndoBERT models. the average loss value obtained from the LSTM model is higher, which makes the accuracy obtained lower than the IndoBERT model.

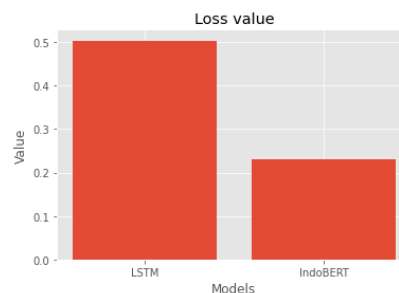


Figure 8. Comparison of Loss Result

#### 4. Conclusion

Based on the results of applying the LSTM and IndoBERT methods using a dataset taken from Twitter about covid-19 for hoax detection tasks. Based on the test results, the LSTM model provides an average accuracy value of 87.54%. And the IndoBERT model from the test results that have been carried out provides an average accuracy value of 92.07%. So from the accuracy results obtained from the LSTM and IndoBERT models, in this study the IndoBERT Model can show good performance in hoax detection tasks and is proven to outperform the LSTM model which can provide the best average accuracy results in this study. Future research can use a larger dataset because the LSTM model requires larger data to perform a task. So that by using a larger dataset, LSTM can provide better performance.

#### References

- [1] "DIGITAL 2022: ANOTHER YEAR OF BUMPER GROWTH - We Are Social UK." <https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/> (accessed May 26, 2022).
- [2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," Aug. 2017, doi: 10.48550/arxiv.1708.01967.
- [3] S. Girgis, E. Amer, and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," *Proceedings - 2018 13th International Conference on Computer Engineering and Systems, ICCES 2018*, pp. 93–97, Feb. 2019, doi: 10.1109/ICCES.2018.8639198.
- [4] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The role of user profiles for fake news detection," *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, pp. 436–439, Aug. 2019, doi: 10.1145/3341161.3342927.
- [5] C. S. Atodiresei, A. Tănăsescu, and A. Iftene, "Identifying Fake News and Fake Users on Twitter," *Procedia Comput Sci*, vol. 126, pp. 451–461, Jan. 2018, doi: 10.1016/J.PROCS.2018.07.279.
- [6] S. K. Dirjen *et al.*, "Hoax Detection System on Twitter using Feed-Forward and Back-Propagation Neural Networks Classification Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 655–663, Aug. 2020, doi: 10.29207/RESTI.V4I4.2038.
- [7] S. Deepak and B. Chitturi, "Deep neural approach to Fake-News identification," *Procedia Comput Sci*, vol. 167, pp. 2236–2243, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.276.
- [8] B. Irena and E. B. Setiawan, "Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 711–716, Aug. 2020, doi: 10.29207/RESTI.V4I4.2125.
- [9] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models," *Procedia Comput Sci*, vol. 179, pp. 704–712, Jan. 2021, doi: 10.1016/J.PROCS.2021.01.059.
- [10] R. K. Kaliyar, K. Fitwe, P. Rajarajeswari, and A. Goswami, "Classification of Hoax/Non-Hoax News Articles on Social Media using an Effective Deep Neural Network," *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, pp. 935–941, Apr. 2021, doi: 10.1109/ICCMC51019.2021.9418282.
- [11] J. Patihullah and E. Winarko, "Hate Speech Detection for Indonesia Tweets Using Word Embedding And Gated Recurrent Unit," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, pp. 43–52, Jan. 2019, doi: 10.22146/IJCCS.40125.
- [12] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, Apr. 2021, doi: 10.1016/J.IJIMEI.2020.100007.
- [13] A. Apriliyanto and R. Kusumaningrum, "HOAX DETECTION IN INDONESIA LANGUAGE USING LONG SHORT-TERM MEMORY MODEL," *SINERGI*, vol. 24, no. 3, pp. 189–196, Jul. 2020, doi: 10.22441/SINERGI.2020.3.003.
- [14] L. Hulliyatus Suadaa, I. Santoso, A. Tabitha, and 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, Jul. 2021, doi: 10.22146/IJCCS.66205.
- [15] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020, doi: 10.48550/arxiv.2009.05387.
- [16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," pp. 757–770, Nov. 2020, doi: 10.48550/arxiv.2011.00677.
- [17] A. M. Abdelhameed, H. G. Daoud, and M. Bayoumi, "Deep Convolutional Bidirectional LSTM Recurrent Neural Network for Epileptic Seizure Detection," *2018 16th IEEE International New Circuits and Systems Conference, NEWCAS 2018*, pp. 139–143, Dec. 2018, doi: 10.1109/NEWCAS.2018.8585542.
- [18] I. R. Hidayat and W. Maharani, "General Depression Detection Analysis Using IndoBERT Method," *International Journal on Information and Communication Technology (IJOICT)*, vol. 8, no. 1, pp. 41–51, Aug. 2022, doi: 10.21108/IJOICT.V8I1.634.
- [19] T. A. S. Rohmah and W. Maharani, "Personality Detection on Twitter Social Media Using IndoBERT Method," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, pp. 448–453, Sep. 2022, doi: 10.47065/BITS.V4I2.1895.
- [20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.
- [21] S. Khomsah and A. S. Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 648–654, Aug. 2020, doi: 10.29207/RESTI.V4I4.2035.