SINERGI Vol. 24, No. 3, October 2020: 189-196
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2020.3.003 .

HOAX DETECTION IN INDONESIA LANGUAGE USING LONG SHORT-TERM MEMORY MODEL

Andi Apriliyanto Retno Kusumaningrum*
Department of Informatics, Universitas Diponegoro
Jl. Prof. Soedarto SH, Tembalang, Semarang 50275, Indonesia
*Corresponding Author Email: retno@live.undip.ac.id

Abstract -- *Nowadays, the internet and social media grow fast. This condition has positive and negative effects on society. They become media to communicate and share information without limitation. However, many people use that easiness to broadcast news or information which do not accurate with the facts and gather people's opinions to get benefits or we called a hoax. Therefore, we need to develop a system that can detect hoax. This research uses the neural network method with Long Short-Term Memory (LSTM) model. The process of the LSTM model to identify hoax has several steps, including dataset collection, pre-processing data, word embedding using pre-trained Word2Vec, built the LSTM model. Detection model performance measurement using precision, recall, and f1-measure matrix. This research results the highest average score of precision is 0.819, recall is 0.809, and f1-measure is 0.807. These results obtained from the combination of the following parameters, i.e., Skip-gram Word2Vec Model Architecture, Hierarchical Softmax, 100 as vector dimension, max pooling, 0.5 as dropout value, and 0.001 of learning rate.*

Keywords: *Hoax Detection; Neural Network; Long Short-Term Memory; Word2vec*

Received: March 12, 2020 *Revised: April 23, 2020* *Accepted: May 1, 2020* *Published: July 15, 2020*

INTRODUCTION

The use of the internet and social media today is familiar. Ease of access and use is an attraction for people to use the internet and social media. Various information and news that are trending can be obtained easily and quickly using the internet and social media, but not all information and news that are spread on the internet and social media that contain facts. Most internet and social media users rely on the latest news [1][2]. This situation will cause devastating news to be widely spread that can lead readers or recipients to negative opinions or often called a hoax.

Hoax is misleading human perception by spreading wrong information, but it is considered as truth [3]. The spreading of hoax news in cyberspace spread deliberately by individuals or groups of people to the community. In 2017 data from the Ministry of Communication and Information Technology (KEMKOMINFO) stated that there were 800 thousand internet sites in Indonesia that were indicated as spreading wrong news and hate speech. Many irresponsible people who use the internet to gain personal and group benefits by spreading negative news that causes anxiety in the community. This condition is getting worse because, according to DailySocial.id research, 44% of Indonesian people cannot detect hoax news.

Various studies on detecting hoax news in Indonesia have been carried out with the use of word similarity measurement theories such as Levenshtein Distance [4]. In addition, research on hoax detection has also been applied using various classical machine learning methods such as K-Nearest Neighbor [5], Decision Tree [6], Naïve Bayes [7], Support Vector Machine [7], Random Forest [8], and so forth [9][10]. However, the application of classical machine learning methods has weaknesses, including not suitable for processing large and complex data, requires an expert to label data and feature extraction manually, cannot learn based on raw data, and difficulty in representing data. The weaknesses of classical machine learning can be overcome by using the application of deep learning methods.

Deep learning has the advantage of studying the hidden representation of the input in both context and content compared to traditional approaches where modeling of extraction features is made manually [11].

However, researches using deep learning methods are still rarely used for cases of detection of Indonesian document hoaxes. Different with English document hoax detection, the use of deep learning methods has been developed with various models and techniques such as Convolutional Neural Network [12][13], Vanilla Recurrent Neural Network [14], Long Short-Term Memory [1], as well as Gated Recurrent Units [15].

Convolutional Neural Network (CNN) is one of deep learning methods that can overcome the shortcomings of classical learning. However, CNN is still unable to process sequential data. Unlike the Recurrent Neural Network (RNN) [16]. RNN is specifically designed to handle subsequent data. However, although it is designed to handle sequential data that work sequentially, RNN has limited capturing long dependencies.

A Short-Term Long Memory (LSTM) model was designed to overcome this limitation. The LSTM model is one variant of the RNN methods. The LSTM model is able to overcome long-term dependencies by remembering long-term information and is good to apply in the case of sentiment analysis or classification, such as hoax news detection. Based on the description above, this study discusses the development of hoax news detection in Indonesian using the LSTM model.

METHOD

This research, in its implementation, has several stages. The stage includes the formation of the dataset, the generation of a hoax detection model, the real-time hoax detection process. At the step of forming the dataset, there are processes of data collection, pre-processing, and word2vec training. The next stage is part of the process of generating a hoax detection model. At this stage, the results of the word2vec dataset model will be carried out a k-fold cross-validation process, which will separate the data into training data and test data for use in the training and testing process of the LSTM model. Then the results will be evaluated to determine the performance of the LSTM model and get the best model that will be used in the real-time hoax detection stage. The general description of the process in this study is shown in Figure 1.

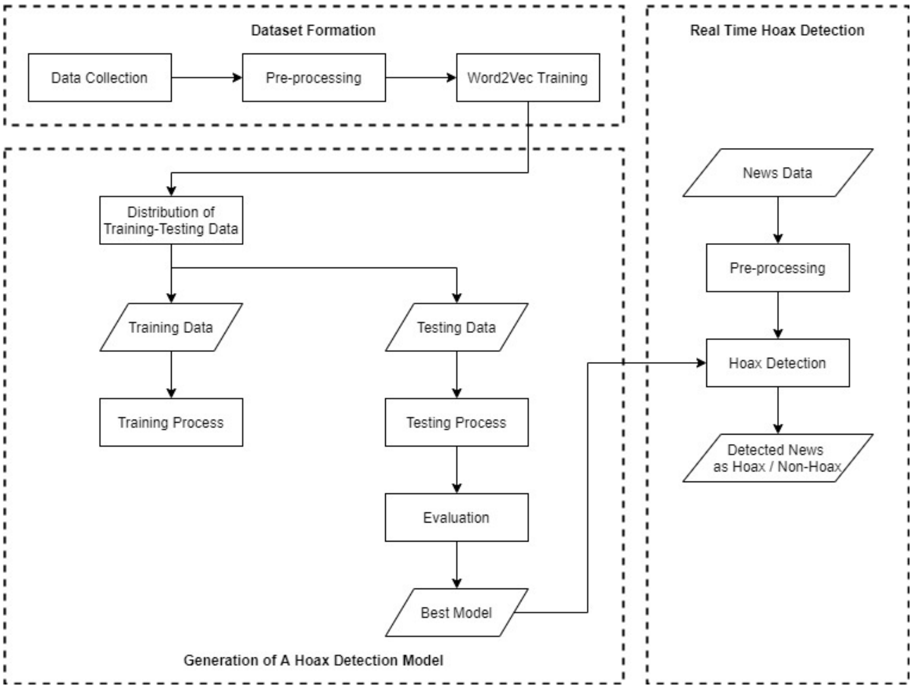

Figure 1. Overview of Research

Data Collection

Data collection in this study uses Indonesian-language news narratives found on the website cekfakta.com, https:// turnbackhoax.id, and cnnindonesia.com. Hoax news is taken from the website cekfakta.com, and https://turnbackhoax.id while non-hoax news is taken from the website cnnindonesia.com.

Pre-processing

Pre-processing is used to process news data into data that can be used in the next process-pre-processing consists of several stages, i.e., case folding, tokenization, filtering, stemming, and padding.

Word2vec Training

The word2vec training process is used to study word vector representations. After the pre-processing process is complete, the dataset that has been processed will be converted into a vector representation so that it can be used in the LSTM algorithm. Word2vec training process is performed based on neural network architecture with a single hidden layer, and the training process aims to learn the weights of the hidden layer, which is subsequently stated as word vectors.

The word2vec training process in this study uses the Python Gensim library. The word2vec model applies 12 combinations by experimenting with the use of two types of word2vec architectural models, the CBOW model and the skip-gram model, two evaluation methods, i.e., Hierarchical Softmax and Negative Sampling, and three dimensions with sizes of 100, 200, and 300.

Distribution of Training Data and Test Data

The process of distributing training data and test data use K-fold cross-validation. The k-value used in this k-fold cross-validation is 10. With 1000 existing news datasets, and using 10-fold cross-validation will produce 100 datasets for each fold. For every 10th iteration, 900 news datasets will be used as training data, and 100 news datasets will be used as test data. For distributing data, it is carried out in a balanced manner each fold by dividing the two classes, i.e., 50 hoax class data and 50 non-hoax class data.

Training and Testing

The training and testing process in this research is divided into two parts of the process, i.e., the training process that aims to build models using LSTM and the testing process to determine the performance of the models that have been built in the training process.

a. Training

The training process uses word arrays that have gone through the word embedding process in the previous stage. The training process uses a combination of parameters as listed in Table 1 to get the right LSTM model.

Table 1. LSTM Model Parameters tested

No.	Parameters	Parameter Type
1.	*Pooling*	Max, dan Average
2.	*Dropout*	0.25; 0.5; 0.75
3.	*Learning Rate*	0.0001; 0.001; 0.01

The LSTM model used in the encoding layer in this study is a model introduced by Hochreiter & Schmidhuberc [17]. The LSTM model is designed to overcome long-term dependencies by remembering long-term information using the gate mechanisms. On the LSTM model, architecture has three gates, i.e., input gate *I*, forget gate *f*, output gate *o*, and a memory cell *c*. LSTM has the ability to add or subtract information into the cell state that is ordered by the gate. The following are the equation in LSTM:

$$i_t = \sigma(W_i x(t) + U_i h(t-1) + b_i) \quad (1)$$

$$f_t = \sigma(W_f x(t) + U_f h(t-1) + b_f) \quad (2)$$

$$o_t = \sigma(W_o x(t) + U_o h(t-1) + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c x(t) + U_c h(t-1) + b_c) \quad (4)$$

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (5)$$

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

As explained in Table 1, to get the most optimal detection model, we modify the LSTM model by applying various values to 3 test parameters, namely pooling type, dropout value, and learning rate value. In addition, because the detection process aims to detect binary classes (hoaxes and non-hoaxes) so that at the fully connected layer, we apply the sigmoid activation function where this function aims to transform input values into ranges of 0.0 - 1.0.

The architecture of the LSTM model can be seen in Figure 2. The input of the LSTM model architecture is an array generated from the word2vec process. Word2vec will change the word order into a vector sequence and then will be forwarded to the LSTM model, and then LSTM will predict the class as an output vector at time step t. LSTM uses c_t cells at each time step t in this LSTM. This process will be iterated to update the hidden state and produce output.

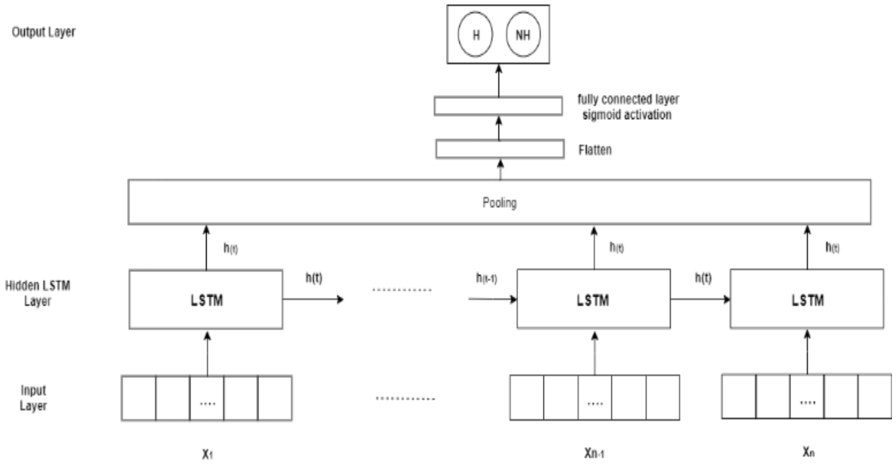

Figure 2. LSTM Model Architecture

b. Testing

This process aims to see the results of the LSTM training process on each combination of parameters. At this stage, an evaluation of the accuracy value is used when using test data. The test results with the highest accuracy value when evaluating are the best models. The best model is used as the LSTM model that will be used in the detection process.

Evaluation

The evaluation technique applied in this research is the confusion matrix. The confusion matrix analyzes how the detection model can recognize tuples from different classes and is also used to measure the performance of the detection model. Table 2 shows the confusion matrix for detection.

Table 2. Example of Confusion Matrix

Confusion Matrix		**Prediction**	
		Positive (+)	*Negative* (-)
Actual	*Positive* (+)	TP	FN
	Negative (-)	FP	TN

where:

- True Positive (TP) if the prediction results are positive, and the data are actually positive.
- True Negative (TN) if the prediction result is negative, and the data is actually negative.
- False Negative (FN) if the result of the prediction is negative, and the data is actually positive.
- False Positive (FP) if the prediction result is positive, and the data is actually negative.

$$Precision = \frac{TP}{FP + TP} \quad (7)$$

$$Recall = \frac{TP}{FN + TP} \quad (8)$$

$$F1 - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (9)$$

RESULTS AND DISCUSSION

Research Data

This research uses a total of 1000 news narrative data. The data division is in the form of 500 news narrative data labeled hoaxes and 500 news narrative data labeled non-hoaxes-news narrative data obtained from the website cekfakta.com, https://turnbackhoax.id, and cnnindonesia.com.

Research Scenarios

In this study, several parameters are used to determine the performance of these parameters on the LSTM model. The parameters in the word2vec model used in this study are the CBOW architecture model and the Skip-gram model. The evaluation method uses Hierarchical Softmax and Negative Sampling. Meanwhile, the dimensions are 100, 200, 300. The word2vec model is then processed using the LSTM model. In the LSTM model, there are parameters in the form of pooling, i.e., Max and Average, dropout parameters with the tested value are 0.25, 0.5, 0.75, and Learning rate parameters with values of 0.0001, 0.001, 0.01. All combined parameters on the word2vec model and parameters of the LSTM model are combined to get the best model. The performance of the parameter combination is determined by the calculation of matrix precision, recall, and f1-measure. The overall value from the obtained combinations will be compared and analyzed to determine the effect of the entered parameters.

The following scenarios will be used in this study:

1) Scenario 1 is done to find out the performance of the word2vec model, i.e., the CBOW architecture model and the Skip-gram model.
2) Scenario 2 is done to find out the performance of the word2vec model, which is an evaluation method using the Hierarchical Softmax and Negative Sampling of the LSTM model.
3) Scenario 3 is done to find out the performance of the word2vec model that is the dimension with a value of 100, 200, 300 to the LSTM model.
4) Scenario 4 is done to find out the pooling performance of the LSTM model. The pooling parameters tested are Max and Average.
5) Scenario 5 is done to find out the dropout performance on the LSTM model. Dropout parameters tested are 0.25, 0.5, and 0.75.
6) Scenario 6 is done to find out the learning rate performance of the LSTM model. The learning rate parameters tested are 0.0001, 0.001, and 0.01.

Research Results and Analysis

a. Scenario 1

In scenario 1 it can be concluded that the architecture of the Skip-gram model produces an average value of precision, recall, and f1-measure better than the CBOW model architecture. The Skip-gram architecture model obtained an average value of precision, recall, and f1-measure of 0.747, 0.750, 0.739, while the CBOW architectural model only obtained an average value of precision, recall, and f1-measure of 0.673, 0,691, 0.669. Skip-gram works better because it predicts a given context from one word, whereas CBOW predicts one word from a context word. This makes Skip-gram superior in predicting unique words that rarely appear compared to CBOW. The results of scenario 1 can be seen in Figure 3.

b. Scenario 2

In scenario 2 it can be concluded that Hierarchical Softmax produces average precision, recall, and f1-measure values better than Negative Sampling. Because the Hierarchical Softmax evaluation method during the training process uses the binary tree model to present all words in the vocabulary and the leaf node contains words that rarely appear so that the word will have the same vector representation as above. Whereas Negative Sampling only updates a sample of some output words as negative samples [18]. The results of scenario 2 can be seen in Figure 4.

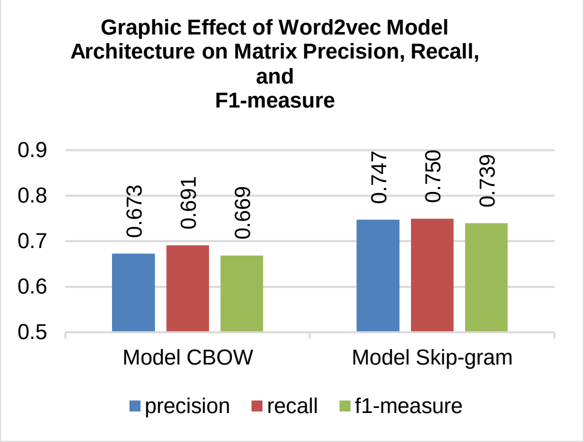

Figure 3. Graphic Effect of Word2vec Model Architecture on Matrix Precision, Recall, and F1-measure

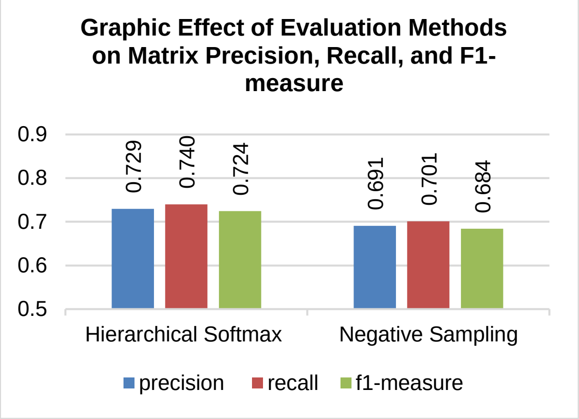

Figure 4. Graphic Effect of Evaluation Methods on Matrix Precision, Recall, and F1-measure

c. Scenario 3

In scenario 3 it can be concluded that the higher the dimension, the lower the average value of matrix precision, recall, and f1-measure. The best dimension value is 100. Increasing the dimension and the amount of data can decrease the accuracy value, so it is necessary to increase the dimension and the amount of data simultaneously [19]. The results of scenario 3 can be seen in Figure 5.

d. Scenario 4

In scenario 4 it can be concluded that max-pooling is higher than average pooling. This situation has happened because max pooling is better used in binary classification problems because it can distinguish nearly equal data with few differences. The results of scenario 4 can be seen in Figure 6.

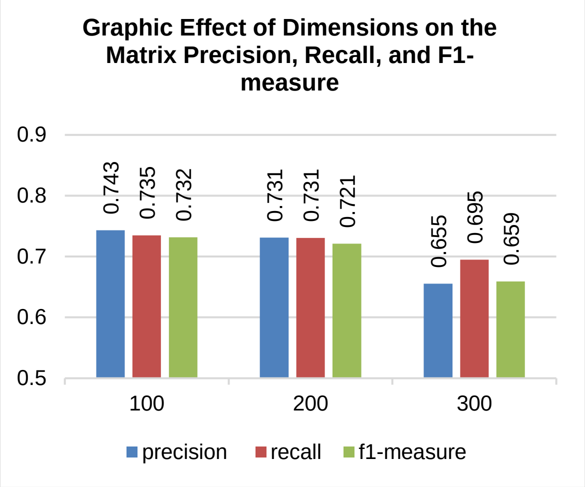

Figure 5. Graphic Effect of Dimensions on Matrix Precision, Recall, and F1-measure

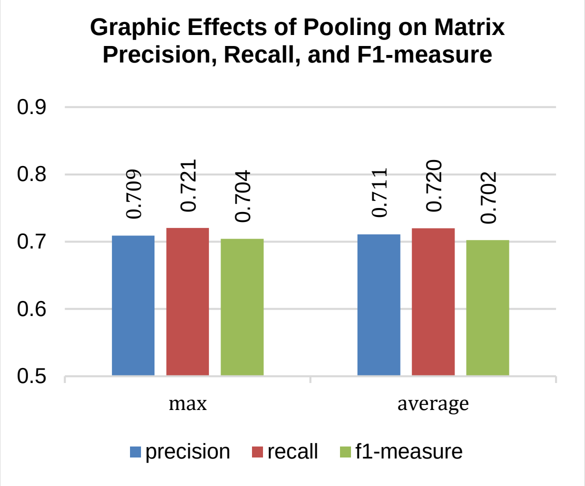

Figure 6. Graphic Effects of Pooling on Matrix Precision, Recall, and F1-measure

e. Scenario 5

In scenario 5 it can be concluded that the dropout with a value of 0.5 has the highest average value compared to 0.25 and 0.75. This is because too small or too high a dropout value will affect the value of the precision matrix, recall, and f1-measure. The dropout value also represents the number of neurons to be removed. The number of neurons affects the learning ability of a model. A large number of datasets also affect the dropout rate, i.e., the greater the dropout value will increase the accuracy of the model in large datasets but will decrease the accuracy in smaller datasets [20]. The results of scenario 5 can be seen in Figure 7.

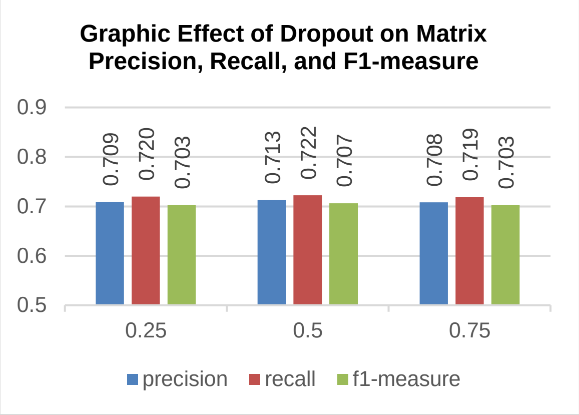

Figure 7. Graphic Effect of Dropout on Matrix Precision, Recall, and F1-measure

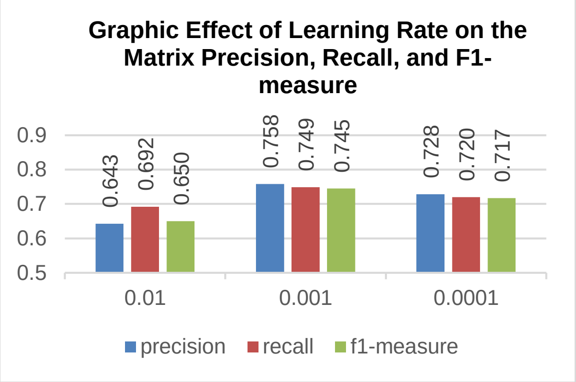

Figure 8. Graphic Effect of Learning Rate on Matrix Precision, Recall, and F1-measure

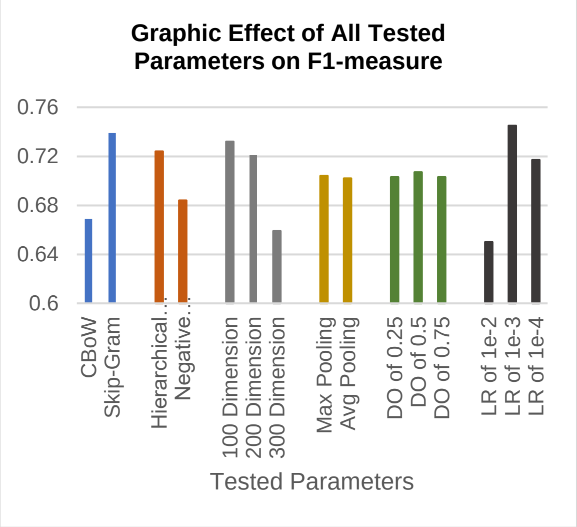

Figure 9. Comparison of All Tested Parameters

f. Scenario 6

In scenario 6 it can be concluded that the best learning rate is at the value of 0.001. This is because the weight changes that occur are not so large and not too small at the learning rate value of 0.001. That way, there is no minimum local missed, and if the learning rate is too large, it will also affect changes in weights, which will make training difficult to converge. The results of scenario 6 can be seen in Figure 8.

Based on Figure 3 to Figure 8, it can be concluded that the best combination parameters, as shown in Figure 9. The best value of each scenario produces a combination of parameters of the word2vec model with the LSTM model. The combination of parameters is the word2vec model Skip-gram architecture model. The evaluation method uses Hierarchical Softmax, dimension 100, LSTM model in the form of Max pooling, dropout parameter with a value of 0.5, and Learning rate parameter with a value of 0.001. The combination produced the average values of the precision, recall, and f1-measure matrices of 0.819, 0.809, and 0.807.

CONCLUSION

The detection of hoax news in the Indonesia language using the LSTM method has been successful. Based on several experiments, it obtained an average value of the precision, recall, and f1-measure matrix of 0.819, 0.809, and 0.807. The highest average value of the confusion matrix is obtained when using the word2vec model parameter combination. The parameter combination contains the Skip-gram architecture model, evaluation method using Hierarchical Softmax, dimension 100 with LSTM model in the form of Max pooling, dropout parameter with value 0.5, and Learning rate parameter with value 0.001.

REFERENCES

[1] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Information Processing & Management*, vol. 57, no. 2, pp. 1-13, March 2020. DOI: 10.1016/j.ipm.2019.02.016

[2] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais," Behind the cues: A benchmark study for fake news detection," *Expert System with Application,* vol. 128, pp. 201-213, August 2019. DOI: 10.1016/j.eswa. 2019.03.036

[3] M. M. Hossain, M. F. Labib, A. S. Rifat, A. K. Das and M. Mukta, "Auto-correction of English to Bengali Transliteration System using Levenshtein Distance," *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, Malaysia, 2019, pp. 1-5. DOI: 10.1109/ICSCC.2019.8843613

[4] R. K. Kaliyar, A. Goswami and P. Narang, "Multiclass Fake News Detection using Ensemble Machine Learning," *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, Tiruchirappalli, India, 2019, pp. 103-107. DOI: 10.1109/ IACC48062.2019.8971579

[5] D. Keskar, S. Palwe, and A. Gupta, "Fake News Classification on Twitter Using Flume, *N*-Gram Analysis, and Decision Tree Machine Learning Technique," In Bhalla S., Kwan P., Bedekar M., Phalnikar R., Sirsikar S. (eds) *Proceeding of International Conference on Computational Science and Applications. Algorithms for Intelligent Systems*. Springer, Singapore, 2020. DOI: 10.1007/978-981-15-0790-8_15

[6] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake News Detection Using Sentiment Analysis," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 2019, pp. 1-5. DOI: 10.1109/IC3.2019.8844880

[7] M. A. Rahmat, Indrabayu and I. S. Areni, "Hoax Web Detection for News in Bahasa Using Support Vector Machine," *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2019, pp. 332-336, DOI: 10.1109/ICOIACT46704.2019. 8938425

[8] M. Dong, L. Yao, X. Wang, B. Benatallah, X. Zhang, and Q. Z. Sheng, "Dual-stream Self-Attentive Random Forest for False Information Detection," *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-8. DOI: 10.1109/IJCNN.2019.8851765.

[9] K. Poddar, G. B. Amali D., and K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, India, 2019, pp. 1-5. DOI: 10.1109/i-PACT44901.2019.8960044

[10] H. S. Al-Ash, M. F. Putri, P. Mursanto, and A. Bustamam, "Ensemble Learning Approach on Indonesian Fake News Classification," *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2019, pp. 1-6. DOI: 10.1109/ICICoS48119.2019.8982409

[11] A. Bondielli, F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences,* vol. 497, pp. 38-55, September 2019. DOI: 10.1016/j.ins. 2019.05.035

[12] Y. C. Chen, Z. -Y. Liu, H. -Y. Kao, "IKM at SemEval-2017 Task 8: Convolutional Neural

Networks for stance detection and rumor verification," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* pp. 465-469, 2017. DOI: 10.18653/v1/S17-2081

[13] D. L. Z. Astuti, S. Samsuryadi, and D. P. Rini, "Real-Time Classification of Facial Expressions using A Principal Component Analysis and Convolutional Neural Network," *SINERGI*, vol. 23, no. 3, pp. 239-244, October 2019. DOI: 10.22441/sinergi. 2019.3.008

[14] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* pp. 797-806, Nov 2017. DOI: 10.1145/3132847.3132877

[15] S. Bajaj, "The Pope Has a New Baby!" Fake News Detection Using Deep Learning," *CS 224N Report*, pp.1-8, 2017

[16] P. T. Tin, "A Study on Deep Learning for Fake News Detection," *JAIST: Japan Advanced Institute of Science and Technology,* pp.1-49, 2018. [Online]. Available: https://dspace.jaist.ac.jp/dspace/bitstream/10119/15196/3/paper.pdf

[17] C. Olah, "Understanding LSTM Networks," *colah.github.io,* 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[18] X. Rong, "word2vec Parameter Learning Explained," *arxiv*, pp. 1-21, 2016. [Online]. Available: https://arxiv.org/abs/1411.2738

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in neural information processing systems,* pp. 1-9, Oct 2013

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research,* vol. 15, no. 56, pp.1929-1958, 2014