Jurnal Teknik Informatika (JUTIF)
Vol. x, No. y, Juni 2021, hlm. x-y

DOI: 10.20884/jutif...
p-ISSN: 2723-3863
e-ISSN: 2723-3871

THE PREDICTION OF RETWEET USING LONG SHORT-TERM MEMORY METHOD WITH THE TOPIC OF COVID-19 VACCINATION

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstract

In last 2019, the Covid-19 outbreak was first reported, infecting at least 20.1 million people and killing more than 737,000 people worldwide and still counting. In Indonesia, the government announced that the Covid-19 vaccination is an obligation for everyone. The rapid development of technology has made social media a platform of spreading news. One of the social media that plays an important role is Twitter. The tweets can be shared with other users by re-tweeting process.The greater the number of retweets, the wider the existing information. Therefore, the retweet feature plays a crucial role in spreading information. This study discusses retweet predictions about Covid-19 vaccination using the Long Short-Term Memory (LSTM) method with the application of hyperparameter tuning to obtain the best result and get an accuracy or closeness values (98%),a precision or closeness value (74%), a recall value (91%), and an f-1 score value or a comparison value (93%).

Keywords: *Covid-19, LSTM, Prediction, Retweet, Twitter, Vaccine*

1. INTRODUCTION

1.1 Backgroundof the Study

The 2019 Covid-19 outbreak was first reported in Wuhan, China in last 2019. It has spread to 216 countries and territories [1]. It has infected at least 20.1 million people and killed more than 737,000 people worldwide and still counting [2], [3]. The Indonesian government has implemented to keep a distance from each other and always wear masks [4], [5]. This strategy can largely prevent and reduce infected community due to the susceptibility of additional waves of infection [4], [6]. According to Health officials, elderly and children are vulnerable to infect it quickly because they have very high health sensitive conditions. It is widely accepted that the world will not return to its pre-pandemic state of normality until a safe and effective vaccine is available [11].

Vaccination is the process of injecting or dripping a vaccine to increase the production of antibodies to prevent certain diseases. Vaccines not only protect themselves, but also provide protection for people who cannot be vaccinated, such as people of a certain age and people with certain diseases. Vaccines do not cause disease. Its used in the community are safe and usually do not cause serious side effects [7]–[9].

Social media is one of the platform for citizens to express their wishes, disseminate information, or get information. One of the social media that plays an important role is Twitter. On Twitter, there is a retweet feature for the users to share such information between other users so that their followers know what information they are receiving[10].

Several machine learning approaches can be used in makinf retweet predictions. Previous studies predicted retweets on Twitter used several classification models, such as K-Nearest Neighbor (KNN), Naive Bayes, Support Vector Machine (SVM) and Random Forest[11].

This study focuses on predicting the spread information on Twitter by predicting whether a tweet will be retweeted or not related to the topic of Covid-19 vaccination using the Long Short-Term Memory (LSTM) classification method. One of the functions of LSTM is that easy to overcome long-term dependence and can determine the value to be selected as the appropriate output on the given input and it has a low value error[12].

The topics discussed in this study are how to get User Based and Content Based features, as well as how to apply the Long Short-Term Memory (LSTM) method in predicting a tweet disseminated. The problem limitation of this final project is that the features used are the User Based and Content Based. The method used is only the Long Short-Term Memory method and the dataset used data from Twitter with an Indonesian-language of Covid-19 vaccination.

2. METHOD

The purpose of this study is to build a model that can predict the spread of information on Twitter using the *Long Short-Term Memory* (LSTM) method which applied several features, such as Total_of_tweet, No_of_followers, No_of_following, Age_of_account,No_of_favourite,aver_favou_per_day,aver_tweets_per_day,user_name_length,contain_

video,contain_picture,contain_user_mention,contain_upper,contain_hastag,contain_rt_sugges,len_of_text, retweeted.

Long Short-Term Memory (LSTM) is a development of the Recurrent Neural Network (RNN) which is able to extract abstract features of data sensor sequence values from the technique manually. In the RNN, there is a simple cell that contains only one layer of neurons with an activation function, such as tanh. In the LSTM architecture, the cell contents are more complex than those of the RNN. The large number of cells in LSTM makes this model can explore the length pattern of time series data because vinishing gradient conditions can be avoided. In other words, LSTM and RNN have differences only in the contents of the cell, but theoretically the principle is same.

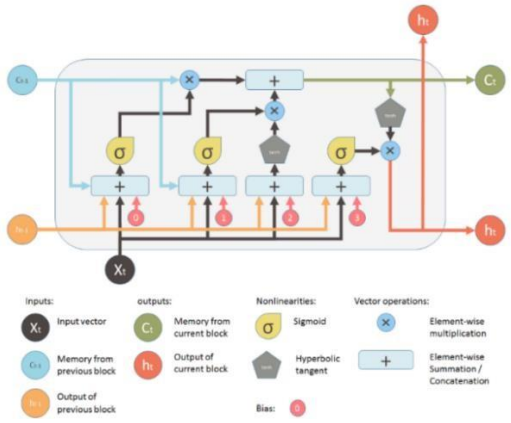

Figure 1. LSTM Architecture

Here are the formula of each gate in LSTM: [13]

1. Forget Gate
   In this process, the input is processed by combining with the previous output value and then executed through the sigmoid activation function. This gate determines whether the previous information is forgotten or not. This information continues in the memory cell or cell state.

   ft = σ (Wf . [xt + ht − 1] + bf) (2.1)
   Information:
   ft: forget gate
   σ: sigmoid activation function
   Wf: weights forget
   gatext: input cell
   HT−1: PreviouslyBF
   cell output: BIAS forget gate

2. Input Gate
   In this process, the previous output value is combined with the current input value. Then, two activation functions must be executed. One path passes through the sigmoid activation function for the input value, the other path passes through the tanh activation function for the candidate memory cell value.

   it = σ (Wi . [x t + ht − 1] + bi)

   Ct~ = tanh(WC. [xt + ht − 1] + bC)

   Information:
   it: input gate Wi: weights input
   WC: weights candidate bC: candidate bias
   Tanh: Tanh Act/ivation function

3. Cell State
   In this process, the two values are combined. The first value is the forget gate value multiplied by the previous cell sate value. The second value is the value of the input gate multiplied by the value of the candidate memory cell.

   Ct = ft × Ct − 1 + it × C˜t (2.4)

   Information:
   Ct: cell stateCt−1: cell state previously

4. Output Gate
   This process produces an output value, where the value comes from combining the previous value with the current value that has gone through the sigmoid activation function.

   ot = σ(Wo × [xt + ht − 1] + bo)(2.5)

   Information:
   ot: gateWo
   output: weights gate
   bo output: bias gate output

5. Hidden Layer
   Hidden Layer affects the value in the next process.The value of this layer is the output value multiplied by the cell state or memory cell value activated by the tanh function.

A Confusion Matrix is a table or matrix containing four values that become a measure of the performance by the classification problem[13]. The four values are presented in the Table 1.

Table 1. Confusion Matrix

		True Values	
		Positive	Negative
Predictions	Positive	TP	FP
	Negative	FN	MR

Description:

TP (True positive): Positive predictions and correct on target.

TN (True negative): Negative prediction and correct on target.

FP (False positive): Positive predictions and false on target.

FN (False negative): Negative Predictions and false not on target.

After the confusion matrix is recognized, it can also be known the accuracy, precision, recall and f1-score values. Here's an explanation and formula to find out: [13]

1. Accuracy

   Accuracy is the calculation of how precise the classification has been built, according to the existing target.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\ \% \quad (2.6)$$

2. Precision

   Precision is the accuracy calculation between the target data and the predicted results of the model.

$$Precision = \frac{TP}{TP + FN} X\ 100\% \quad (2.7)$$

3. Recall

   Recall is a calculation that describes the success of a model in rediscovering information.

$$Recall = \frac{TP}{TP + FN} \text{ x } 100\% \quad (2.8)$$

4. F1-Score

   F1-score is a calculation that describes the comparison between precision and recall. If the FN and FP values are not close to the f1- score, it should be used instead of the accuracy value.

$$F1 - Score = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} x\ 100\% \quad (2.9)$$

The system design that built in this study is a system that can predict retweets of Covid-19 vaccination into retweet classes and non-retweet classes. Formed flowchart according to the order are as follows.

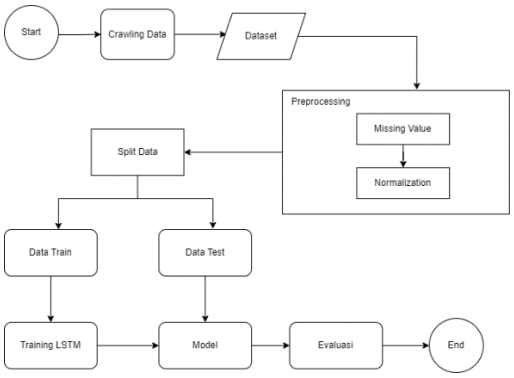

Figure 2. *Flowchart System*

Based on Figure 2, the system flow is started by collecting Twitter data. The data is saved into excel form. Furthermore, the data will be processed consisting of *check missing value* and *normalization.* Next, split data are carried out, and training will later be used to train algorithms in finding the appropriate model, while testing data are used to test and find out the performance of the model obtained at the testing stage. The data that has been tested by the model then produce predictions and also system evaluations.

Is a stage of the data collection process. The dataset used in this study is crawled data sourced from Twitter. Data collection used Netlytic by searching for Indonesian tweets with the keyword "Covid-19 vaccination". The dataset obtained amounted to 11719 with 12 user-based and content-based features used.

- *User_statuses_count* : number of tweets/stsatus that users have uploaded
- *user_followers_count* : number of followers of the user account
- *user_friends_count* : number of user friends
- *age_of_account* : user's age
- *favorite_count* : number of users' favorites
- *aver_favou_per_day* : the average obtained from the division between *user_followers_count* and *age_of_account*
- *aver_tweets_per_day* : the average obtained from the division between *youser_statuses_count* and *age_of_account*
- *user_name_length* : length of user account name
- *contain_user_mention* : Tweets uploaded mentioning other users
- *contain_hashtag* : tweets that use hashtags
- *len_of_text* : length of a user's tweet
- *retweet_count* : number of retweets

Data pre-processing is the next step after data collection. It aims to process data that will be input into the system[10]. The several stages were performed as follow:

1) Data Cleaning

   Ensure that there are no missing values or empty data and eliminate duplicate data.

2) Case Folding
To normalize the data using the Min-Max Normalization method. It works by converting each existing feature into a range value between 0 and 1. The equation is as follows.

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}} \quad (3.1)$$

Before classifying, data sharing or data split was is carried out with a comparison ratio of 80:20, 80% for the data train and 20% for the test data. Then, LSTM Modeling was performed after the feature extraction is completed. This stage proceed structured data with time steps. Each time step consists of several gates, namely forget gate, input gate, cell state, output gate used to calculate the output value of the hidden layer value [13].

Hyperparameter Tuning is a stage to produce an optimal model and improve the performance of a model adapted to a particular model [14]. In this study, Hyperparameter was obtained by *the grid search* method. It is a method to try combinations of several hyperparameters. Then, compared each combination with the same calculation metric. The experiment was conducted in two values, namely the learning rate value and the batch size. Batch size is one of the hyperparameters that affect the performance of a neural network. This happens because the batch size can indirectly affect the amount of weight contained in the LSTM architecture.

This study used 11 datasets that had been collected, 719 with 12 types of features. The dataset used in this process has previously been preprocessed using check missing value and normalization. After that, data sharing or split data is carried out with a comparison ratio of 80:20, which means 80% of the data train and 20% of the test data. The feature used to predict retweets is 'retweetcount'. Evaluation in this study was carried out in several testing scenarios, namely testing data using min-max normalization and without the application of min-max normalization to find out the best scenario that produced the most optimal performance value. Then, to find out the optimal performance results, there were values from the evaluation results in the form of accuracy, precision, recall, and f1-score values.

3. RESEARCH RESULT

In scenario 1 testing the dataset uses *min-max normalization* and uses *hyperparameter tuning* in the *Long Short-Term Memory* classification model. Testing in scenario 1 to find out whether data using min-max normalization performed better than datasets without *min-max normalization* applied.

Table 2. Scenario 1 Test Results

Preprocessing	Accuracy	*Precision*	*Recall*	*F1-Score*
Min-max normalization	0.83	0.83	1.00	0.91

Based on Table 2, it was found that the accuracy value using min-max normalization was 83%. Determining the proper use of preprocessing can affect the level of performance of the model used.

In scenario 2, dataset testing was performed to view without the application of min-max normalization and using hyperparameter tuning in the Long Short-Term Memory classification model. Testing in scenario 2 to find out whether data without min-max normalization implementation performed better than datasets using min-max normalization.

Table 3. Scenario 2 Test Results

Preprocessing	Accuracy	*Precision*	*Recall*	*F1-Score*
Without *Min-max normalization*	0.98	0.96	0.91	0.93

Based on Table 2, it was found that the accuracy value using min-max normalization is 83%.

According to the results of the tests, several scenarios have been performed issued different performance results. The first test results in Table 2 show that a prediction model built on a predetermined feature the normalization process does not have much effect on the existing data. Besides, in subsequent tests, testing with datasets was carried out without going through the min-max normalization stage. From the results of the first and the second scenario where data without the application of min-max normalization obtained accuracy results of 98%, precision values of 74%, recall values of 91%, f-1 scores of 93%, and for preprocessing using min-max normalization, the accuracy result is 83%, the precision value is 83%, the recall value is 100%, and the f-1score is 91%.

There are several previous studies that have been carried out as literature reviews in making the final project. Research conducted by Hoang T and Mothe J in [5] predicted the diffusion of information on Twitter using *multi-class* classifications [15] with the types of features used in modeling of user-based, time-based, and content-based. Through the result of F-measure performance for both binary and multi-class prediction types, it increased by 5% and the most important feature was the number of followers and user groups. The time-based feature is highly correlated to the retweet ability.

In addition, research conducted by Molaei, Soheila, Hadi Zare, and Hadi Veisi in [16] discussed the learning approach of new meta-path representations about Heterogeneous Deep Diffusion (HDD) and predicted information diffusion by applying CNN-LSTM to generates a representation of metapaths. DBLP, PubMed, and ACM are the data

sets used in studies related to the results that all three datasets show the effective and efficient LSTM and CNN-LSTM methods.

Other researches have been conducted by Syeda Firdaus, Chen Ding, and Alireza Sadegian in [17] which provides an overview of research in the field of retweet prediction that can be used as a guide for the future research.The features used in this study are user-based, author-based, and content-based. The retweet prediction model used the extracted features. Meanwhile the dataset used the Twitter API. Through the results despite numerous research efforts on retweet predictions, the accuracy of the predictions was far from perfect. There is a major challenge for the researchers in finding reasons and methods to effectively disseminate information through online social networks.

Online social networks are gaining popularity as information channels, especially Twitter. Twitter provides a variety of information ranging from health, education, sports, and politics. The information disseminated through the social network Twitter is a challenge for researchers, because the available information can maximize the impact of its spread, so as to use it more effectively. It not only helps to study it, but can also be used to solve the problem of disseminating information on the social network Twitter [18].

4. CONCLUSION

In this study, intending to build a retweet prediction system using the Long Short-Term Memory method, the performance value showed quite a good result. This can be seen from the performance results with the application of hyperparameter tuning without min-max normalization in getting the best value with an accuracy or proximity value of 98%, precision value or proximity value of 74%, recall value of 91%, f-1 score or a comparison value of 93%. The higher value of the accuracy, the more optimal the prediction rate. For the data tested using min-max normalization, the user does not have much effect on the dataset used in this study. For further research, new features can be developed with other classification and preprocessing methods to determine which method can produce the best accuracy.

Bibliography

[1] M. Jeyanathan, S. Afkhami, F. Smaill, M. S. Miller, B. D. Lichty, and Z. Xing, "Immunological considerations for COVID-19 vaccine strategies," *Nat. Rev. Immunol.*, vol. 20, no. 10, pp. 615–632, 2020, doi: 10.1038/s41577-020-00434-6.

[2] G. Singh, A. S. Aiyub, T. Greig, S. Naidu, A. Sewak, and S. Sharma, "Exploring panic buying behavior during the COVID-19 pandemic: a developing country perspective," *Int. J. Emerg. Mark.*, 2021, doi: 10.1108/IJOEM-03-2021-0308.

[3] P. T. Leeson and H. A. Thompson, "Public choice and public health," *Public Choice*, 2021, doi: 10.1007/s11127-021-00900-2.

[4] D. A. Rantauni and E. Sukmawati, "Correlation of Knowledge and Compliance of Implementing 5m Health Protocols in the Post-Covid-19 Pandemic Period," Online, 2022. [Online]. Available: www.midwifery.iocspublisher.orgjournalhomepage:www.midwifery.iocspublisher.org

[5] B. Hu, S. Huang, and L. Yin, "The cytokine storm and COVID-19," *Journal of Medical Virology*, vol. 93, no. 1. 2021. doi: 10.1002/jmv.26232.

[6] Y. C. Wu, C. S. Chen, and Y. J. Chan, "The outbreak of COVID-19: An overview," *Journal of the Chinese Medical Association*, vol. 83, no. 3. 2020. doi: 10.1097/JCMA.0000000000000270.

[7] I. Iskak, M. Z. Rusydi, R. Hutauruk, S. Chakim, and W. R. Ahmad, "Meningkatkan Kesadaran Masyarakat Tentang Pentingnya Vaksinasi Di Masjid Al – Ikhlas, Jakarta Barat," *J. PADMA Pengabdi. Dharma Masy.*, vol. 1, no. 3, 2021, doi: 10.32493/jpdm.v1i3.11431.

[8] UNICEF, "Vaksin COVID-19 & KIPI," *Unicef*, 2021.

[9] Ellyzabeth Sukmawati, Norif Didik Nur Imanah, and Dahlia Arief Rantauni, "PENGETAHUAN IBU TENTANG VAKSIN COVID-19 DENGAN MOTIVASI IBU UNTUK MEMBERIKAN VAKSIN PADA ANAK YANG DI SEKOLAH DASAR," *Forikes Forum Ilm. Kesehat.*, vol. 13, 2023, Accessed: Oct. 20, 2022. [Online]. Available: http://forikes-ejournal.com/ojs-2.4.6/index.php/SF/article/view/2361

[10] D. Febrianto and K. Muslim Lhaksmana, "Prediksi Retweet Dengan Fitur Berbasis Pengguna dan Tingkat Sentimen Menggunakan Metode Klasifikasi Naive Bayes," vol. 8, no. 5, pp. 11200–11206, 2021.

[11] M. S. Zannuar and K. M. Lhaksmana, "Prediksi Retweet Berdasarkan Feature User-Based Menggunakan Metode Klasifikasi Random Forest," vol. 8, no. 5, pp. 11183–11191, 2021.

[12] W. Hastomo, A. Satyo, and Sudjiran, "Long short term memory machine learning untuk memprediksi akurasi nilai tukar IDR terhadap USD," *J. SeNTIK*, vol. 3, no. 2, pp. 115–124, 2019.

[13] J. Nurvania and K. M. Lhaksamana, "Analisis Sentimen Pada Ulasan di TripAdvisor Menggunakan Metode Long Short-Term Memory (LSTM)," *e-Proceeding Eng.*, vol. 8, no. 4, pp. 4124–4135, 2021.

[14] J. Nurvania, Jondri, and K. Muslim Lhaksamana, "Analisis Sentimen Pada Ulasan di TripAdvisor Menggunakan Metode Long Short-Term Memory (LSTM)," *e-Proceeding Eng.*, vol. 8, no. 4, pp. 4124–4135, 2021.

[15] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictive features," *J. Comput. Sci.*, vol. 28, pp. 257–264, 2018, doi: 10.1016/j.jocs.2017.10.010.

[16] S. Molaei, H. Zare, and H. Veisi, "Deep learning approach on information diffusion in heterogeneous networks," *Knowledge-Based Syst.*, vol. 189, 2020, doi: 10.1016/j.knosys.2019.105153.

[17] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet: A popular information diffusion mechanism – A survey paper," *Online Soc. Networks Media*, vol. 6, pp. 26–40, 2018, doi: 10.1016/j.osnem.2018.04.001.

[18] I. P. Dewi, J. Jondri, and K. M. Lhaksmana, "Prediksi Retweet Menggunakan Metode Bernoulli Dan Gaussian Naive Bayes Di Media Sosial Twitter Dengan Topik Vaksinasi Covid-19," *eProceedings Eng.*, vol. 8, no. 5, pp. 11216–11225, 2021.