JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 2 (2023)
https://doi.org/10.35877/454RI.jinav1864

Classification of Multi-Label of Hate Speech on Twitter Indonesia using LSTM and BiLSTM Method

Elita Aurora Az Zahra, Yuliant Sibaroni, & Sri Suryani Prasetyowati

Faculty of Informatics, Universitas Telkom, Bandung, Indonesia

Abstract

Social media is a communication tool that supports users to interact socially using technology. One of the most popular social media platforms is Twitter. However, its media platform has been considered by the virtual police as one of the main sources of spreading hate speech on social media. In this final project research, the authors conducted a study on the detection of hate speech in tweets on Twitter Indonesia. The method used in this research is multi-label classification by applying the LSTM and BiLSTM methods. The dataset used was 13,169 tweet data, and data labeling process was carried out into 12 classes. The results revealed that the LSTM and BiLSTM methods had good performance in classifying text data with 10 trials with an accuracy value of 78.67% for LSTM and 80.25% for BiLSTM. Based on the accuracy obtained, BiLSTM has higher accuracy than LSTM, so it can be concluded that BiLSTM is superior to LSTM.

Keywords: twitter, hate speech, social media, LSTM, BiLSTM

Received: 27 June 2023 Revised: 3 October 2023 Accepted: 23 October 2023

1. Introduction

Social media is a communication tool that can support the social interaction of its users by using web-based technology to become an interactive dialogue (Watie, 2016). Users can participate in social media to provide feedback and comments openly and for an unlimited time (Watie, 2016). Social media is also used as a means of disseminating information in various fields, such as business, tourism, education, politics, health, and so on. However, in the dissemination of such information, there are quite few who misuse information on social media (Cahyono, 2016). The government has also issued a Law on Information and Electronic Transactions contained in Article 27 paragraph 3 of the ITE Law (Lindawati et al., 2021; Putri et al., 2015; Terkini, 2022).

Twitter is one of the most popular social media in Indonesia with as many as 18.45 million users in Indonesia in early 2022 (We Are Social, 2022). Twitter is considered by the virtual police as one of the social media that contributes a lot to the spread of hate speech (Al Ayyubi, 2021; Cardaioli et al., 2020; Liu et al., 2021; Zubiaga et al., 2016). According to the Head of the Public Information Section of the National Police, *Kombes Pol* Ahmad Ramadhan, in the period 23 February - 12 April 2021, there were 195 Twitter accounts that were caught in the virtual police due to tweets containing hate speech and ethnicity, religion, race, and intergroup relations *(SARA)* (Al Ayyubi, 2021). It refers to the hate speech as an act of communication in the form of words, actions, or actions carried out by one person or group with the aim of provoking and insulting another person or group in various aspects, such as race, skin color, ethnicity, religion, and so on (Sasongko et al., 2021).

Based on these problems, an action is needed such as detecting hate speech on Twitter Indonesia. There are several studies that detect hate speech on Indonesian Twitter. One of them is research by (Fadli & Hidayatullah, 2021) which shows that this research can identify cyberbullying on social media Twitter Indonesia using the LSTM and BiLSTM methods. The accuracy value obtained in this study was 93.77% for LSTM and 95.24% for BiLSTM. However, this study only divides the data into two classes, namely the cyberbullying class and the non-cyberbullying class. The research used a multi-label classification using the BiLSTM method (Ilma et al., 2021). This study carried out multi-

* Corresponding author.
E-mail address: elitaaurora@student.telkomuniversity.ac.id

label data labeling into 12 labels of hate speech. This study obtained good performance result with an accuracy value of 82.31%.

In this study, the authors conducted research on hate speech detection using multi-label classification and comparing the LSTM and BiLSTM methods. This study used a method that has been used in previous studies (Fadli & Hidayatullah, 2021), namely using LSTM and BiLSTM. The reason is that the LSTM and BiLSTM algorithms were quite good at classifying text data. In labelling process, the data involved a multi-label classification, namely entering data into several categories in the hope of increasing the success of this study.

Research on the detection of hate speech was carried out by (Hidayatullah, 2019) using the Support Vector Machine and Multinominal Naïve Bayes algorithms on Indonesian tweets. This study employed tweet data by utilizing the Twitter API and then chopped into two classes, namely rude tweets and not-rude tweets. This study compares two models for classification, namely Support Vector Machine with an accuracy of 0.9928, a precision of 0.9914, a recall of 0.9946, and an F-1 score of 0.9930. Meanwhile, Multinominal Naïve Bayes produced an accuracy of 0.9943, a precision of 0.9912, a recall of 0.9762, and an F-1 score of 0.9836. The results of this study indicate that the Support Vector Machine is superior in classifying data. However, the difference in values for the two algorithms is not much different, so it can be concluded that the two algorithms have almost the same performance in classifying text.

Other research on the detection of cyberbullying on Indonesian Twitter (Fadli & Hidayatullah, 2021) was carried out using the deep learning method with the LSTM and BiLSTM algorithms. The data taken is in the form of tweets on Twitter Indonesia. Then, data labeling was carried out into two classes, namely the cyberbullying class and the non-cyberbullying class. The results of the evaluation of this study indicate that the BiLSTM algorithm is superior in classifying tweet data compared to LSTM. The value of LSTM is accuracy of 93.77, precision of 91.59, recall of 92.45, and F1-Score of 92.02 from LSTM. While the values of BiLSTM are accuracy of 95.24, precision of 94.29, recall of 93.40, and F-1 Score of 93.84. The results of the values from LSTM and BiLSTM have a slight difference in values so that it can also be proven that LSTM is also a pretty good algorithm for classifying tweet data.

Research on multi-label classification on Indonesian Twitter social media has been conducted by (Dwitama & Hidayat, 2021) using a deep learning algorithm with the Convolution Neural Network method. In this study, the model is search using several test parameters, such as learning rate, number of convolution layers, convolution kernel size, number of *nth* layers, and number of nodes in *nth* layer, and supported using BERT tokenizer. At the end of the test, a comparison of model development was carried out by applying class weighting and without applying class weighting. The results obtained the model development that is applied using class weighting is lower than the model development without applying class weighting with an accuracy value of model development using class weighting of 98.07% and without class weighting of 98.76%.

Research on multi-label classification has also been carried out by (Ilma et al., 2021) using the Bidirectional Long Short-Term Memory (BiLSTM) method. This research labeled each tweet that had been collected by Twitter crawling into 12 labels about hate speech. This study obtained good performance, namely accuracy value of 82.31%, precision of 83.41%, recall of 87.28%, and F-1 score of 85.30%.

Based on these studies, the author tried to develop detection of hate speech on Indonesian Twitter. Multi-label classification using the LSTM and BiLSTM methods has never been studied before. This research was conducted using the method used by Habib Faizal Fadli and Ahmad Fathan Hidayatullah (Fadli & Hidayatullah, 2021) as a reference by developing data labeling into several classes using multi-label classification.

(Hidayatullah, 2019) using the Support Vector Machine and Multinominal Naïve Bayes algorithms on Indonesian tweets. This study used tweet data by utilizing the Twitter API and then chopped into two classes, namely rude tweets and not rude tweets. This study compares two models for classification, namely Support Vector Machine with an accuracy of 0.9928, a precision of 0.9914, a recall of 0.9946, and an F-1 score of 0.9930. Meanwhile, Multinominal Naïve Bayes produced an accuracy of 0.9943, a precision of 0.9912, a recall of 0.9762, and an F-1 score of 0.9836. The results of this study indicate that the Support Vector Machine is superior in classifying data. However, the difference in values for the two algorithms is not much different. Thus, it can be concluded that the two algorithms have almost the same performance in classifying text.

Other research on cyberbullying detection on Indonesian Twitter by (Fadli & Hidayatullah, 2021) carried out using the deep learning method with the LSTM and BiLSTM algorithms. The data taken is in the form of tweets on Twitter Indonesia. Then, data labeling was carried out into two classes, namely the cyberbullying class and the non-cyberbullying class. The results of the evaluation of this study indicate that the BiLSTM algorithm is superior in classifying tweet data compared to LSTM. The value of LSTM is accuracy of 93.77, precision of 91.59, recall of 92.45,

and F1-Score of 92.02 from LSTM. Meanwhile, the values of BiLSTM are accuracy of 95.24, precision of 94.29, recall of 93.40, and F-1 Score of 93.84. The results of the values from LSTM and BiLSTM have a slight difference in values so that it can also be proven that LSTM is also a pretty good algorithm for classifying tweet data.

Research on multi-label classification on Indonesian Twitter social media has been conducted (Dwitama & Hidayat, 2021) using deep learning algorithms with the Convolution Neural Network method. In this study, the model was searched using several test parameters, such as learning rate, number of convolution layers, convolution kernel size, number of nth layers, and number of nodes in nth layer and supported using BERT tokenizer. Then, at the end of the test, a comparison of model development is carried out by applying class weighting and without applying class weighting. The results obtained showed that the model development applied using class weighting is lower than the model development without class weighting application, with an accuracy value of model development using class weighting of 98.07% and without class weighting of 98.76%.

Research on multi-label classification has also been carried out by (Ilma et al., 2021) using the Bidirectional Long Short-Term Memory (BiLSTM) method. This research labeled each tweet that has been collected by Twitter crawling into 12 labels about hate speech. This study obtained good performance, namely accuracy value of 82.31%, precision of 83.41%, recall of 87.28%, and F-1 score of 85.30%.

Based on these research studies, the author will try to develop detection of hate speech on Indonesian Twitter. Multi-label classification used the LSTM and BiLSTM methods has never been studied before. This research was conducted using the method by Habib Faizal Fadli and Ahmad Fathan Hidayatullah (Fadli & Hidayatullah, 2021) as a reference by developing data labeling into several classes using multi-label classification.

2. Research Methods

The purpose of this research aims to detect hate speech on Indonesian Twitter using LSTM and BiLSTM, to find out the performance of multi-label classification on Indonesian Twitter hate speech tweets using the LSTM and BiLSTM methods and to compare the LSTM and BiLSTM methods on hate speech tweets on Indonesian Twitter using multi-label classification.

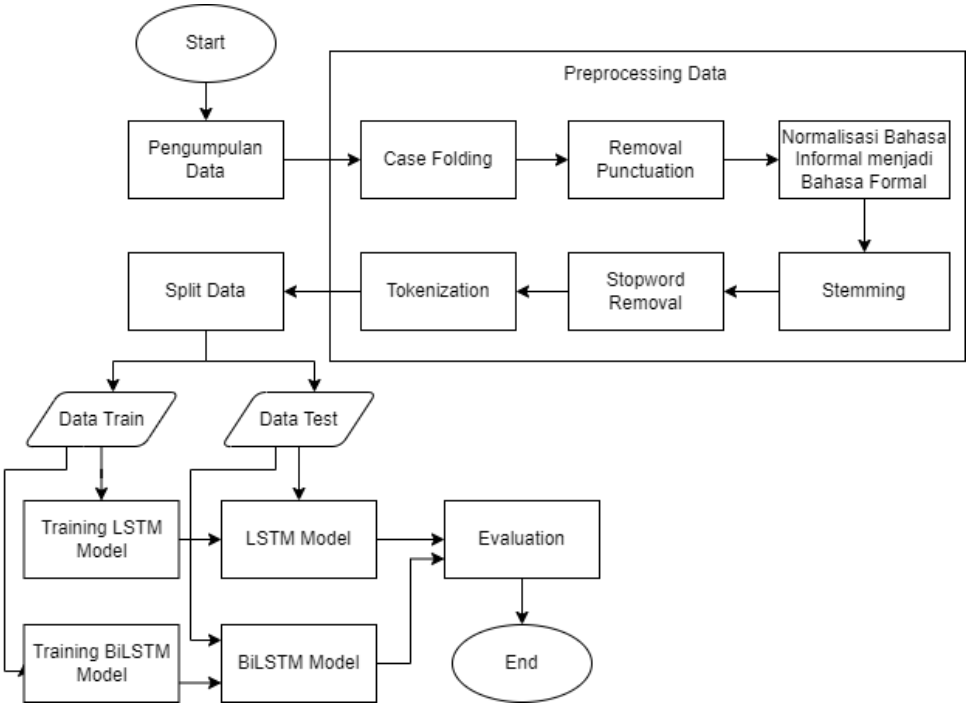

Figure 1. System Design

The system design in this research involved data collection using the available data on the Kaggle platform. The data is already in a multi-label format with 12 classes, as shown in Table 3. Next, data was then carried out with the initial stages, namely case folding, followed by punctuation removal, normalization of informal language into formal language,

stemming, stopword removal, and finally tokenization. After the data preprocessing completed, the data were split into two, namely train data and test data, with a split data of 80% train data and 20% test data. In the data train, LSTM model training and BiLSTM model training was carried out. After the training was completed, next to LSTM Model and BiLSTM Model process. The LSTM and BiLSTM Model were used for test data. After that, next to an evaluation process using a confusion matrix. The data collection process was carried out using data already available on the Kaggle platform, and the data has been collected and multi-labeled into 12 classes (Ibrohim & Budi, 2019).

3. Results and Discussion

In these data, there are 13.169 data tweets from Twitter Indonesia. The multi-label labeling is divided into 12 classes as shown in Table 2.

Table 2. Data Labeling Categories

No.	Category	Description
1.	Hate Speech (HS)	Hate speech class.
2.	Abusive	Class of offensive language.
3.	HS_Individual	Hate speech directed by an individual.
4.	HS_Group	Hate speech directed towards a group of people.
5.	HS_Religion	Hate speech related to religion.
6.	HS_Race	Hate speech related to race.
7.	HS_Physical	Hate speech related to physical appearance.
8.	HS_Gender	Hate speech related to specific gender or sexual orientation.
9.	HS_Other	Hate speech unrelated to religion, race, physical appearance, or gender.
10.	HS_Weak	Hate speech that does not contain provocation or incitement.
11.	HS_Moderate	Hate speech that only occurs within social media.
12.	HS_Strong	Hate speech directed at an individual or group that provokes or incites.

The collected data during the data collection process was then carried out by preprocessing the data. It is more effective to process so that during the classification process and can produce more accurate values. The following steps are carried out in preprocessing:

a) Folding Case

The first stage in preprocessing this data is the process of case folding, which aims to change all letters to lowercase with the letters 'A' to 'Z' changed to the letters 'a' to 'z'. Case folding was done using the Sastrawi python library by adding the '.lower()' function.

b) Removal Punctuation

The next step is removal punctuation, namely removing characters, such as '!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~' in the tweet data using the literary python library.

c) Normalization of Informal Language into Formal Language

After the punctuation removal completed, the informal language was normalized into a formal language, such as *'asyikl'* to *'asyik'* by using a data dictionary taken from Github.

d) Stemming

Next, stemming was carried out by removing affixes such as '*-nya'* in the data. Stemming was done using import StemmerFactory in the Sastrawi python library.

e) Stopword Removal

Then, the stopword removal process was carried out, namely the removal of common words such as *'ada', 'dan' 'untuk', 'dari', 'di', 'maka', 'atau'*, and so on. Stopword removal was done using a stopword data dictionary taken from Github.

f) Tokenization

Then tokenization was carried out by breaking sentences on each data into word lists. Tokenization was done using word_tokenize() in the NLTK module.

The preprocessing results can be seen in Table 3.

Table 3. Preprocessing Results

Input	Preprocessing	Output
USER Nah.!!! Itu die..\nWaktu thn 1965, anggota dewan yg bangga jadi anak pki, usianya brp ya?? '	*Case Folding*	*nah.!!! itu die.. waktu thn , anggota dewan yg bangga jadi anak pki, usianya brp ya??'*
nah.!!! itu die.. waktu thn , anggota dewan yg bangga jadi anak pki, usianya brp ya??'	Removal Punctuation	*nah itu die waktu thn anggota dewan yg bangga jadi anak pki usianya brp ya*
nah itu die waktu thn anggota dewan yg bangga jadi anak pki usianya brp ya	Normalization of Informal Language into Formal Language	*nah itu dia waktu tahun anggota dewan yang bangga jadi anak partai komunis indonesia usianya berapa ya*
nah itu dia waktu tahun anggota dewan yang bangga jadi anak partai komunis indonesia usianya berapa ya	Stemming	*nah itu dia waktu tahun anggota dewan yang bangga jadi anak partai komunis indonesia usia berapa ya*
nah itu dia waktu tahun anggota dewan yang bangga jadi anak partai komunis indonesia usia berapa ya	Stopword Removal	*anggota dewan bangga anak partai komunis indonesia usia ya*
anggota dewan bangga anak partai komunis indonesia usia ya	Tokenization	*'anggota', 'dewan', 'bangga', 'anak', 'partai', 'komunis', 'indonesia', 'usia', 'ya'*

Long Short-Term Memory (LSTM) is a modification of the Recurrent Neural Network (RNN) because RNN has limitations in remembering old data. Thus, an LSTM was formed that is able to remember data stored for a long time (Staudemeyer & Morris, 2019). In the LSTM architecture, there are three gates (input gate, output gate, forget gate), and a memory cell. The cell remembers a value for each time interval and the three gates regulate the flow of information in and out of the cell. At each time step, the LSTM obtained the input from the current time step and the output from the previous time step, as well as produced an output that was fed to the next time step. The hidden layer from the last step was called the hidden layer used as a classification (Minaee et al., 2019).

In this study, LSTM used *'tf.keras.Sequential()'* with the model consisting layer embedding, dropout, LSTM, and dense using activation ReLU and softmax. The first layer is LTSM architecture model of embedding to take the input word sequence and return the embedding vector for each word in the sequence. The embedding layer in this study used vocab_size = 50000 which determined the size of the vocabulary, embedding *dim* as the resulting vector embedding dimension, and input length using the maximum length of the text to be used. Next, the dropout layer was used to avoid overfitting with a dropout level of 0.3 or 30%. Next, the LSTM layer with 64 units. Then, a dense layer with 64 units and used ReLU activation. This layer was used to connect each LSTM unit to the next units. The next layer is a dense layer with the number of units equal to the number of labels in the study, namely 12 with softmax activation. This layer was used to generate output. Next, the model was compiled using the loss function = 'categorical_crossentropy', the optimizer used is 'adam', and the observed metric is accuracy. Furthermore, training process was carried out on the LSTM model using the '.fit()' method on the train data with a number of iterations of 10 and 20. During the training process, an evaluation was carried out on the LSTM model with test data to see its performance.

Bidirectional Long Short Term Memory (BiLSTM) is a continuation of Long Short Term memory (LSTM) by connecting two hidden layers of input on BiLSTM, namely the forward input used to represent previous information and the backward input used to represent later information which was then incorporated into the architecture from BiLSTM, sourced from the opposite direction to the same output (Isnain et al., 2020; Rizky et al., 2021). Through this form of deep learning, layers of neurons could simultaneously obtain information from past and future conditions (Isnain et al., 2020). Bidirectional Long Short Term Memory (BiLSTM) architecture is presented in Figure 2.

After the LSTM and BilSTM classification processes have been successfully carried out, the next step is measurement to determine performance by looking for accuracy, precision, recall, and F-1 Score values (Shultz et al., 2011). The value of TP1, …, FN2 can be seen in the confusion matrix table (Table 4).

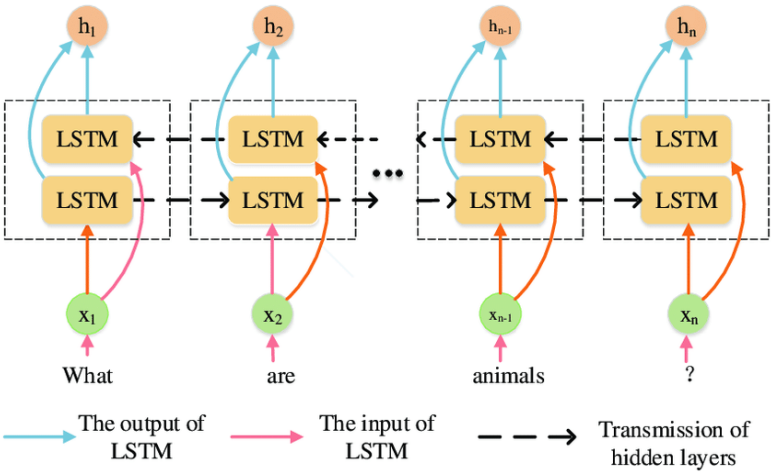

Figure 2. Architecture of Bidirectional Long Short Term Memory

Source: https://www.researchgate.net/figure/Structure-of-bidirectional-long-short-term-memory-LSTM_fig3_343981315

Table 4. Confusion Matrix

Predicted Class	Actual Class	
	True	False
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Posivite)	TN (True Negative)

Data that has passed preprocessing are balanced data. Thus, the dataset has a balanced number of samples. Balance data was performed by under-sampling. Then, the data is divided into 80% train data and 20% test data. All models in this study used the k-fold cross validation algorithm with k = 10 and run using epochs of 10 and 20 with a batch size of 64.

Table 5. Results of the LTSM testing with epoch 10

Column	Precision	Recall	F1-Score	Accuracy
HS	79%	79%	79%	79%
Abusive	87%	87%	87%	87%
HS Individual	74%	74%	74%	74%
HS Group	74%	74%	74%	74%
HS Religion	82%	82%	82%	82%
HS Race	90%	90%	90%	90%
HS Physical	77%	77%	77%	77%
HS Gender	76%	76%	76%	76%
HS Other	79%	79%	79%	79%
HS Weak	71%	71%	71%	71%
HS Moderate	71%	71%	71%	71%
HS Strong	84%	84%	84%	84%
Averagr	**78.67%**	**78.67%**	**78.67%**	**78.67%**

In Table 5 and Table 6, it can be seen that the test results on the LSTM model at epoch 10 obtained an accuracy of 78.67% and at epoch 20 obtained an accuracy of 78,08%. These can be seen that the highest accuracy is found in epoch 10.

Table 7 and Table 8 are the test results of the BiLSTM model. Epoch 10 shows the average accuracy of 80,25% and epoch 20 of 79,75%. It can be seen that the highest accuracy is found in epoch 10.

Table 6. Results of the LTSM testing with epoch 20

Column	Precision	Recall	F1-Score	Accuracy
HS	78%	78%	78%	78%
Abusive	86%	86%	86%	86%
HS Individual	73%	73%	73%	73%
HS Group	73%	73%	73%	73%
HS Religion	81%	81%	81%	81%
HS Race	90%	90%	90%	90%
HS Physical	78%	78%	78%	78%
HS Gender	76%	76%	76%	76%
HS Other	78%	78%	78%	78%
HS Weak	70%	70%	70%	70%
HS Moderate	70%	70%	70%	70%
HS Strong	84%	84%	84%	84%
Average	**78.08%**	**78.08%**	**78.08%**	**78.08%**

Table 7. Results of the LTSM testing with epoch 10

Column	Precision	Recall	F1-Score	Accuracy
HS	97%	97%	97%	97%
Abusive	87%	87%	87%	87%
HS Individual	73%	73%	73%	73%
HS Group	74%	74%	74%	74%
HS Religion	84%	84%	84%	84%
HS Race	89%	89%	89%	89%
HS Physical	77%	76%	76%	76%
HS Gender	76%	76%	76%	76%
HS Other	79%	79%	79%	79%
HS Weak	71%	71%	71%	71%
HS Moderate	72%	72%	72%	72%
HS Strong	85%	85%	85%	85%
Average	**80.33%**	**80.25%**	**80.25%**	**80.25%**

Table 8. Results of the LTSM testing with epoch 20

Column	Precision	Recall	F1-Score	Accuracy
HS	97%	97%	97%	97%
Abusive	86%	86%	86%	86%
HS Individual	73%	73%	73%	73%
HS Group	73%	73%	73%	73%
HS Religion	82%	82%	82%	82%
HS Race	90%	90%	90%	90%
HS Physical	78%	78%	78%	78%
HS Gender	76%	76%	76%	76%
HS Other	78%	78%	78%	78%
HS Weak	70%	70%	70%	70%
HS Moderate	70%	70%	70%	70%
HS Strong	84%	84%	84%	84%
Average	**79.75%**	**79.75%**	**79.75%**	**79.75%**

Table 9 and Table 10 show that the LSTM and BiLSTM modeling have succeeded in detecting hate speech and its categories in tweets using accuracy testing from the experimental results. Tests on the LSTM model were carried out with the parameter vocab_size=50000, embedding_dim=32, input_length=240, dropout=0.3, LSTM=64, and dense=64. In the LSTM model, the parameters vocab_size=50000, embedding_dim=32, input_length=240, dropout=0.3, BiLSTM=64, and dense=64. Then, undersampling was done for the dataset so that it could be balanced. Furthermore, the division of the dataset size was carried out by 80:20. Both models were trained by comparing epoch=10 and epoch=20. From the test results, it was found that accuracy with epoch=10 is higher than epoch=20, because a higher

epoch caused overfitting. A higher epoch makes the model memorize the train data too much so that it cannot generalize well to the test data. At epoch=10 and epoch=20, the high accuracy model is BiLSTM. The BiLSTM model at epoch=10 produces an accuracy of 80.25%. In this study, BiLSTM consistently has higher accuracy than the LSTM in the dataset used. This shows that BiLSTM has the ability to access context information from two directions, providing an advantage in classifying texts, especially when the context of the two directions has a significant influence on understanding the text. BiLSTM is also capable of extracting more complex and diverse features by considering information from two directions. Meanwhile, the LSTM only accesses context information from one direction only.

Table 9. Prediction of LSTM Model

	Tweet	
Column	*Aktor huruhara 98 Prabowo S ingin lengserkan pemerintahan Jokowi … Nyata*	*USER Kasur mana enak kunyuk'*
HS	Positive	Negative
Abusive	Positive	Positive
HS Individual	Negative	Negative
HS Group	Positive	Negative
HS Religion	Positive	Negative
HS Race	Negative	Negative
HS Physical	Negative	Negative
HS Gender	Negative	Negative
HS Other	Negative	Negative
HS Weak	Negative	Negative
HS Moderate	Positive	Negative
HS Strong	Negative	Negative

Table 10. Prediction of LSTM Model

	Tweet	
Column	*Aktor huruhara 98 Prabowo S ingin lengserkan pemerintahan Jokowi … Nyata*	*USER Kasur mana enak kunyuk'*
HS	Positive	Negative
Abusive	Positive	Positive
HS Individual	Negative	Negative
HS Group	Positive	Negative
HS Religion	Positive	Negative
HS Race	Negative	Negative
HS Physical	Negative	Negative
HS Gender	Negative	Negative
HS Other	Negative	Negative
HS Weak	Negative	Negative
HS Moderate	Positive	Negative
HS Strong	Negative	Negative

4. Conclusion

This research succeeded in classifying 12 labels on Indonesian Twitter tweets using LSTM and BiLSTM. The dataset in this study contains 12 hate speech category labels. In the LSTM and BiLSTM modeling processes, model training was carried out for 10 epochs and 20 epochs. The modeling results are better when using 10 epochs with an accuracy of 78.67% for LSTM and 80.25% for BiLSTM. Whereas, 20 epochs resulted in an accuracy of 78.08% for LSTM and 79.75% for BiLSTM. Based on the accuracy results, BiLSTM is consistently superior to LSTM at epoch 10 and epoch 20.

The researcher suggests for further research to explore models other than LSTM and BiLSTM to carry out further performance comparisons between these models to provide more comprehensive insights in selecting the best model for the classification of hate speech. Furthermore, tuning the hyperparameters model can be the good choice so that optimal combinations of hyperparameters can be found to improve model performance.

References

Al Ayyubi, S. (2021). Polri: Ujaran kebencian dan SARA paling banyak di Twitter dan Facebook. *Kabar 24*.

Cahyono, A. S. (2016). Pengaruh media sosial terhadap perubahan sosial masyarakat di Indonesia. *Jurnal Ilmu Sosial & Ilmu Politik Diterbitkan Oleh Fakultas Ilmu Sosial & Politik, Universitas Tulungagung*, *9*(1).

Cardaioli, M., Cecconello, S., Conti, M., Pajola, L., & Turrin, F. (2020). Fake News Spreaders Profiling through Behavioural Analysis Notebook for PAN at CLEF 2020. *CEUR Workshop Proceedings*, *2696*(September), 22–25.

Dwitama, A. P. J., & Hidayat, S. (2021). Identifikasi Ujaran Kebencian Multilabel Pada Teks Twitter Berbahasa Indonesia Menggunakan Convolution Neural Network. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *3*(2), 117. https://doi.org/10.30865/json.v3i2.3610

Fadli, H., & Hidayatullah, A. (2021). Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM. *Universitas Islam Indonesia (UII)*, *2*(No. 1), 1–6.

Hidayatullah, A. F. dkk. (2019). Identifikasi konten kasar pada tweet bahasa Indonesia. *Jurnal Linguistik Komputasional*, *2*(1), 1–5.

Ibrohim, M. O., & Budi, I. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*. 46–57. https://doi.org/10.18653/v1/w19-3506

Ilma, R. A., Hadi, S., & Helen, A. (2021). Twitter's Hate Speech Multi-label Classification Using Bidirectional Long Short-term Memory (BiLSTM) Method. *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, 93–99. https://doi.org/10.1109/ICAIBDA53487.2021.9689767

Isnain, A. R., Sihabuddin, A., & Suyanto, Y. (2020). Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *14*(2), 169. https://doi.org/10.22146/ijccs.51743

Lindawati, Y. I., Setyoningrum, T., & Kunci, K. (2021). Relevansi Penggunaan Media Sosial dengan Hasil Belajar Kognitif Siswa Sekolah Menengah Atas Informasi Artikel ABSTRAK. *Jurnal Ilmiah Kependidikan*, *8*(2).

Liu, R., Gupta, S., & Patel, P. (2021). The Application of the Principles of Responsible AI on Social Media Marketing for Digital Health. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10191-z

Minaee, S., Azimi, E., & Abdolrashidi, A. (2019). *Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models*.

Putri, F. I., Lukmantoro, T., Sos, S., Si, M., Dwiningtyas, H., Ma, S., Joyo, D., & Gono, N. S. (2015). Teknik-teknik Persuasif Dalam Media Sosial. *Jurnal Ilmu Komunikasi*, *3*(1).

Rizky, M. G., Jusak, J., & Puspasari, I. (2021). ANALISIS PERBANDINGAN METODE LSTM DAN BiLSTM UNTUK KLASIFIKASI SINYAL JANTUNG PHONOCARDIOGRAM. *Journal JCONES*, *10*(2), 44–49.

Sasongko, Artanti, V. A. A., Putri, N. U., Hendrawan, J., & Sari, S. D. (2021). Ujaran Kebencian di Media Sosial dalam Perspektif Cyberlaw di Indonesia. *Proceeding of Conference on Law and Social Studies*, 1–12.

Shultz, T. R., Fahlman, S. E., Craw, S., Andritsos, P., Tsaparas, P., Silva, R., Drummond, C., Ling, C. X., Sheng, V. S., Drummond, C., Lanzi, P. L., Gama, J., Wiegand, R. P., Sen, P., Namata, G., Bilgic, M., Getoor, L., He, J., Jain, S., … Mueen, A. (2011). Confusion Matrix. *Encyclopedia of Machine Learning*, 209–209. https://doi.org/10.1007/978-0-387-30164-8_157

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks*.

Terkini, B. (2022). *Bunyi UU ITE Pasal 27 Ayat 3 dan Ancaman Hukumannya | kumparan.com*.

Watie, E. D. S. (2016). Komunikasi dan Media Sosial (Communications and Social Media). *Jurnal The Messenger*, *3*(2). https://doi.org/10.26623/themessenger.v3i2.270

We Are Social. (2022). Digital 2022 Indonesia. *Databoks Katadata*.

Zubiaga, A., Liakata, M., & Procter, R. (2016). *Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media*.