# Bidirectional LSTM for Analyzing Sentiment towards Hate Speech on Twitter.

**Adrianus Charlie Hadirria Agelsadewa1, Hari Suparwito2**

1Informatics Department, Sanata Dharma University, Depok Sleman Yogyakarta, Indonesia, 55281

e-mail: 1adrianus.charlie02@gmail.com

2Informatics Department, Sanata Dharma University, Depok Sleman Yogyakarta, Indonesia, 55281

e-mail: 2hsupsj.usd@gmail.com

|  |  |
| --- | --- |
| Submitted Date: MMMM dd, yyyy | Reviewed Date: MMMM dd, yyyy |
| Revised Date: MMMM dd, yyyy | Accepted Date: MMMM dd, yyyy |

## Abstract

With the rapid development of information technology in Indonesia's entry into the industry 4.0 era, there has been an increase in the use of hate speech on social media, which can trigger division in society. One method to detect tweets containing hate speech is sentiment analysis using deep learning techniques such as Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) algorithms, which can remember information in the long term, remove irrelevant information, and better understand the context of data. This method is essential for social media users to avoid using hate speech and ensure that their tweets do not contain elements of hate speech. This research focused on analyzing the sentiment of Twitter data related to Selena Gomez with a dataset of 11,646 tweets. The study focused on the LSTM and Bi-LSTM model architectures and found that the BiLSTM model with relu activation function for the hidden layer and softmax for the output layer, Adam optimizer, and categorical crossentropy loss function achieved the best and optimal performance with an accuracy of 85.80%.

Keywords: Classification; LSTM and Bi-LSTM; Machine Learning; Sentiment Analysis; Twitter

## Introduction

In Indonesia, the development of information and technology is rapidly growing. Industry 4.0 has already begun in Indonesia, where everything is becoming digital. Twitter has around 18.45 million users in Indonesia and is a constant source of information connecting users with stories, ideas, opinions, and the latest news (Rizaty, 2022). However, many tweets contain hate speech, including those related to the feud between Selena Gomez, Kylie Jenner, and Hailey Bieber, international artists who significantly influence teenagers and adults worldwide. Hate speech can cause division on social media and potentially manipulate public opinion. Therefore, there is a method to detect hate speech using sentiment analysis with LSTM and BiLSTM algorithms. Sentiment analysis allows us to predict whether a tweet contains hate speech based on a word in a sentence. The Bi-LSTM method works bidirectionally in storing information from the past to the future and from the future to the past. Using the Bi-LSTM method, tweets containing hate speech can be predicted more accurately and optimally, particularly in textual data. This study aims to analyze the sentiment of tweets containing hate speech using deep learning with the Bi-LSTM algorithm. Similar studies on sentiment analysis have been conducted to analyze sentiment in textual data, such as sentiment analysis on Twitter using the support vector machine (SVM) approach, which produced good accuracy (Ahmad, Ali, & Ahtab, 2017) (Huq, Ali, & Rahman, 2017). Another study compared the decision tree, Naïve Bayes, and random forest algorithms for sentiment analysis on Twitter, showing that sentiment analysis can be performed using all three algorithms and produces good accuracy, particularly with the Naïve Bayes algorithm (Fitri, Andreswari, & Hasibuan, 2019). A study on sentiment analysis using a deep convolutional network on Twitter showed that sentiment analysis could be done using deep learning algorithms, notably the convolutional neural network algorithm, resulting in good accuracy (Jianqiang, Xiaolin, & Xuejun, 2018) (Ain, et al., 2017). Another study used the Bidirectional GRU (Gated Recurrent Unit) algorithm to analyze sentiment in Arabic, which showed good accuracy and demonstrated that deep learning algorithms, notably the GRU algorithm, can be used bidirectionally (Abdelgwad, Soliman, Taloba, & Farghaly, 2021) (ALI, YANG, QIU, KE, & WANG, 2021). Based on this background and the studies mentioned above, the current study applies the Bidirectional LSTM algorithm to a deep learning model to classify the sentiment of data containing hate speech or not from tweets containing the word "Selena Gomez."

## Methodology

This research will go through several stages, including starting with data collection. Once the data is collected, it will be preprocessed to optimize its quality. Next, the data will be labelled. Then, the text data will be converted into a collection of numerical tokens that represent each word. After being converted into tokens, the dataset will be split into two parts: training and testing dataset. Once the data is ready, two models consisting of layers with LSTM and BiLSTM architecture will be created.

Image 1. The flow of research methodology

A picture containing text, diagram, plan, technical drawing

Description automatically generated

1. **Collecting Data**

This step aims to coolecting data from Twitter platforms using web scraping method. The data that will be collected have three attributes like created date, username and text for one tweet.

1. **Preprocessing Data**

This step aims to structure the text data and improve its quality by removing/transforming some words that are not significant. The following processes will be performed: lowercasing, removing unnecessary words, removing non-alphanumeric characters, stopword removal, and lemmatization.

1. **Labelling Data**

After the data goes through preprocessing, it needs to be labelled because sentiment analysis involves classification, which requires a label for each data in the dataset to build a classification model. Therefore, the data will go through the labelling stage using Vader Lexicon, one of the popular libraries that helps label data in sentiment analysis. Vader requires text input and returns a compound value, which measures data sentiment. The data will be labelled negative if the compound value is less than -0.5. The data will be labelled neutral if the value is between -0.5 and 0.5. Otherwise, it will be labelled positive.

1. **Tokenizing Data**

In this stage, data transformation will occur, where text data needs to be converted into a sequence of numerical tokens. This is necessary because the model being built accepts input in the form of numerical sequences. In this stage, an index of words will be created, where all unique words in the dataset will be converted into a sequence of numerical values. Then, each text data in the dataset will be converted into a sequence of numbers. After this, the data is ready to move on to the modelling stage.

1. **Split Data**

At this stage, the tokenizing dataset will be divided into training and testing datasets, with a ratio of 0.8 for training data and 0.2 for test data.

1. **Modelling**

In this stage, two deep learning models will be constructed with different layer architectures, namely LSTM and BiLSTM. Both models will be built using the same parameters and treatments, including the relu activation function for the hidden layer, softmax activation function for the output layer, adam optimizer, and categorical cross-entropy loss function. The models will then be trained on the training data for 20 epochs with a validation split of 0.1. Additionally, an early stopping callback will be incorporated into the models, which will stop the training process when there is no improvement in accuracy on the validation set.

1. **Evaluation**

The TensorFlow library's evaluation function will be used to evaluate the model. This involves using the test data to test the performance of the trained model. The evaluation metrics used are loss, accuracy, and F1 score. These metrics will comprehensively evaluate the model's performance and ability to classify sentiment in the test data accurately. Using these metrics, we can determine if the model is effective and has room for improvement.

## Results and Discussion

1. **Collecting Data**

At this stage, 11,645 tweet data were obtained with three attributes, including the date the tweet was made, the username of the tweet maker and the tweet text. The following table is an example of scraping data.

Table 1.Example of collected Data

|  |  |  |
| --- | --- | --- |
| created\_at | username | text |
| Thu Mar 02 23:59:30 +0000 2023 | darkonsun | social media is a disease and this hailey bieber selena gomez tea is too hot |
| Thu Mar 02 23:56:18 +0000 2023 | babyyyannieee | Justin Bieber &amp; Selena Gomez followed each other on Twitter 😭teenage me is crying 😭 |

1. **Preprocessing Data**

At this stage, preprocessing will be carried out on the data that has been collected. Data that has been collected will be selected first, which will filter data containing the word "RT" to be removed from the dataset because in tweets, the word "RT" is a representation of duplicate tweet data. Duplication data should be eliminated as it will affect the quality of the model's exercises. After the duplicate data was removed, a dataset with a total of 5206 was obtained.

* + 1. **Lowercasing**

At this stage, all words in the tweet will be transformed into lowercase characters. The following is an example of data from the implementation of lowercasing.

Table 2. Example of lowercasing output

|  |  |
| --- | --- |
| Before | After |
| Social media is a disease and this hailey bieber selena gomez tea is too hot | social media is a disease and this hailey bieber selena gomez tea is too hot |
| Justin bieber &amp; selena gomez followed each other on twitter 😭teenage me is crying 😭 | justin bieber &amp; selena gomez followed each other on twitter 😭teenage me is crying 😭 |

* + 1. **Remove Unnecessary Words**

This step will remove unique characters, such as using "\n" which represents a newline and links. The following is an example of data from the implementation of removing unnecessary words.

Table 3. Example of removing unnecessary words

|  |  |
| --- | --- |
| Before | After |
| social media is a disease and this hailey bieber selena gomez tea is too hot | social media is a disease and this hailey bieber selena gomez tea is too hot |
| justin bieber &amp; selena gomez followed each other on twitter 😭teenage me is crying 😭 | justin bieber &amp; selena gomez followed each other on twitter 😭teenage me is crying 😭 |

* + 1. **Remove Non-Alphanumeric Characters**

This step will remove characters other than letters and numbers in the dataset. For example, the phrase "Hello, world!" would be changed to "Hello world". The following is an example of data from the implementation of removing non-alphanumeric characters.

Table 4.Example of removing non-alphanumeric characters

|  |  |
| --- | --- |
| Before | After |
| social media is a disease and this hailey bieber selena gomez tea is too hot | social media is a disease and this hailey bieber selena gomez tea is too hot |
| justin bieber &amp; selena gomez followed each other on twitter 😭teenage me is crying 😭 | justin bieber amp selena gomez followed each other on twitter teenage me is crying |

* + 1. **Stopword Removal**

This stage serves to eliminate common words, such as conjunctions in language. The following is an example of data from the implementation of stopword removal.

Table 5. Example of stopword removal

|  |  |
| --- | --- |
| Before | After |
| social media is a disease and this hailey bieber selena gomez tea is too hot | social media disease hailey bieber selena gomez tea hot |
| justin bieber amp selena gomez followed each other on twitter teenage me is crying | justin bieber amp selena gomez followed twitter teenage crying |

* + 1. **Lemmatization**

Lemmatization also changes the word into its basic form. Lemmatization ensures that the root word belongs to the language. The following is an example of data from the implementation of lemmatization.

Table 6. Example of lemmatization

|  |  |
| --- | --- |
| Before | After |
| social media is a disease and this hailey bieber selena gomez tea is too hot | social media disease hailey bieber selena gomez tea hot |
| justin bieber amp selena gomez followed each other on twitter teenage me is crying | justin bieber amp selena gomez followed twitter teenage crying |

1. **Labelling Data**

Each existing tweet data will be labelled. Data will be labelled into three categories: neutral, positive and negative. This process uses Vader lexicon sentiment analysis to get a label in each tweet. After this process, the number of data variations was obtained: tweets with positive labels as much as 2091, neutral as many as 1828, and negative as many as 1287. Here are some examples of data after going through the labelling stage.

Table 7. Example of labelling

|  |  |
| --- | --- |
| Text | Label |
| social media disease hailey bieber selena gomez tea hot | Positive |
| justin bieber amp selena gomez followed twitter teenage crying | Negative |

1. **Tokenizing Data**

At this stage, it will convert text data into a collection of tokens in the form of numbers. This is done in order to be able to perform computations and calculations in the model training stage. Each word will be converted to a number based on the index of words created, and the length of the text/token set will be equal by adding the 0 token element at the end of the sequence. Here is an example of a tokenizing implementation on a dataset.

Table 8. Example of Tokenizing

|  |  |
| --- | --- |
| Text | Token |
| social media disease hailey bieber selena gomez tea hot | [ 199, 231, 1737, 4, 5, 2, 3, 222, , 0, 0, 0, 0, 0, 0] |
| justin bieber amp selena gomez followed twitter teenage crying | [ 10, 5, 17, 2, 3, 29, 238, 1090, 365, 0, 0, 0, 0, 0, 0] |

1. **Split Data**

At this stage, the data will be divided into 4164 training data or equivalent to a ratio of 0.8 and test data of 1042 or equivalent to a ratio of 0.2.

1. **Modelling**

Two design models with different architectural layers will be made. The model was built based on LSTM and BiLSTM architectures. The activation function, loss function, and training scheme are the same as in the previous section.

Table 9.The architecture of the LSTM model

|  |  |
| --- | --- |
| Layer | Layer |
| Embedding (128 unit) | Embedding (128 unit) |
| LSTM (128 unit) | BiLSTM (128 unit) |
| LSTM (128 unit) | BiLSTM (128 unit) |
| LSTM (128 unit) | BiLSTM (128 unit) |
| Flatten | Flatten |
| Dense (512 unit) | Dense (512 unit) |
| Dropout (0.5) | Dropout (0.5) |
| Dense (3 unit) | Dense (3 unit) |

1. **Evaluation**

After the modeling stage, the model is trained in accordance with the provisions that have been applied at the beginning. After the training and evaluation stages, the results can be seen in the following table.

Table 10. Results of evaluating each model after the training step

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Loss | F1 Score |
| LSTM | 82.53 % | 0.6372 | 82.52% |
| BiLSTM | 85.80 % | 0.4557 | 85.22% |

Based on the evaluation of the models above, the BiLSTM model with relu activation function on the hidden layer, softmax activation function on the output layer, adam optimizer, and categorical cross-entropy loss function achieved the highest accuracy, lowest loss, and best F1 score. This indicates that the BiLSTM model is a strong candidate for predicting the target variable and can be used confidently for further analysis and applications.

Image 2. LSTM and BiLSTM accuracy

A picture containing text, screenshot, diagram, line

Description automatically generated

The above graph indicates that the BiLSTM model outperforms the LSTM model in terms of accuracy during training. However, both models still experience significant overfitting, as evidenced by the substantial gap between the training accuracy and validation accuracy results. Despite this, the accuracy score of 85.80% during the evaluation stage demonstrates that the BiLSTM model is precise enough in predicting hate speech. Therefore, it can be concluded that the BiLSTM model is a reliable option for further analysis and implementation in natural language processing.

Image 3. LSTM and BiLSTM F1 Score

A picture containing text, screenshot, diagram, line

Description automatically generated

The BiLSTM model outperforms the LSTM model in terms of the F1 score during the training process. However, both models still experience a relatively lower level of overfitting compared to the accuracy. The F1 score of 85.22% during the evaluation stage indicates that the BiLSTM model performs well in classifying data into the correct categories. F1 score is a metric that combines precision and recall, resulting in a score between 0 and 1, where a higher score indicates better model performance. In conclusion, the BiLSTM model is a promising option for further analysis and implementation in natural language processing based on its superior F1 score and lower level of overfitting.

Image 4. LSTM and BiLSTM Loss

A picture containing text, line, screenshot, diagram

Description automatically generated

Image 4 shows that the BiLSTM model performs better in the training process compared to the LSTM model in terms of loss. However, both models still experience overfitting that increases as the epoch increases. The evaluation loss of 0.4557 indicates that the BiLSTM model has a relatively low chance of error in classifying hate speech tweets. The BiLSTM architecture outperforms the LSTM architecture in all three-evaluation metrics. This is because BiLSTM solves the problem of long-term dependency, like LSTM addresses the problem of context in a text (Zhou, Long, & Ou). For example, the same word with different contexts, such as "I like apple fruit" and "I like apple product". BiLSTM can recognize the context of "apple" in these sentences because the model works in two directions, forward and backward. In the Selena Gomez phenomenon, which has spread widely on Twitter, there are various tweet contexts sent by users from different countries with different backgrounds. The BiLSTM layer is suitable and better for classifying hate speech.

## Conclusion

The deep learning approach using LSTM and BiLSTM algorithms successfully analyzed the sentiment of Twitter hate speech on tweets containing the keyword "Selena Gomez". Both LSTM and BiLSTM models demonstrated a reasonably good performance in classifying Twitter hate speech using the same parameters during testing. However, the BiLSTM model outperformed the LSTM model in terms of accuracy, loss, and F1 score due to its ability to recognize the context of a sentence by working forward and backward. The BiLSTM model achieved an accuracy of 85.80%, a loss of 0.4557, and an F1 score of 85.22%, demonstrating its superiority in predicting hate speech.

## References

Abdelgwad, M. M., Soliman, T. H., Taloba, A. I., & Farghaly, M. F. (2021). Arabic aspect based sentiment analysis using bidirectional GRU based models. Retrieved from https://www.sciencedirect.com/science/article/pii/S1319157821002482

Ahmad, M., Ali, I., & Ahtab, S. (2017). *Sentiment Analysis of Tweets using SVM*. Retrieved from https://www.researchgate.net/profile/Shabib-Aftab-2/publication/321084834\_Sentiment\_Analysis\_of\_Tweets\_using\_SVM/links/5a1497b90f7e9b925cd514b0/Sentiment-Analysis-of-Tweets-using-SVM.pdf

Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment Analysis Using DeepLearning Techniques: A Review. Retrieved from https://www.researchgate.net/profile/Amna-Noureen/publication/318096420\_Sentiment\_Analysis\_Using\_Deep\_Learning\_Techniques\_A\_Review/links/596509014585157fcc5e3595/Sentiment-Analysis-Using-Deep-Learning-Te

ALI, W., YANG, Y., QIU, X., KE, Y., & WANG, Y. (2021). Aspect-Level Sentiment Analysis Based on Bidirectional-GRU in SIoT. Retrieved from https://ieeexplore.ieee.org/abstract/document/9424566

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). *Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm*. Retrieved from https://www.researchgate.net/profile/Muhammad-Fauzi-6/publication/327060733\_Random\_Forest\_Approach\_for\_Sentiment\_Analysis\_in\_Indonesian\_Language/links/5d305ff3458515c11c39adfd/Random-Forest-Approach-for-Sentiment-Analysis-in-Indonesian-Language.pdf

Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. Retrieved from https://pdfs.semanticscholar.org/05a8/78000170abcd0c6f8208080470858422e17c.pdf

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). *Deep Convolution Neural Networks for Twitter Sentiment Analysis*. Retrieved from https://ieeexplore.ieee.org/abstract/document/8244338

Rizaty, M. A. (2022). Pengguna Twitter di Indonesia Capai 18,45 Juta pada 2022. Retrieved from https://dataindonesia.id/digital/detail/pengguna-twitter-di-indonesia-capai-1845-juta-pada-2022

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. *European Journal of Operational Research, 218*(1), 211-229. doi:10.1016/j.ejor.2011.09.031

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington: Morgan Kaufmann.

Yap, B. W., Rani, K. A., Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).* *285*, pp. 13-22. Singapore: Springer. doi:10.1007/978-981-4585-18-7\_2

Zhou, K., Long, F., & Ou, W. (n.d.). Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention. Retrieved from https://ieeexplore.ieee.org/abstract/document/8845615