BAREKENG: Journal of Mathematics and Its Applications
June 2023 Volume 17 Issue 2 Page 1113-1122
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss2pp1113-1122

TEXT CLASSIFICATION OF TWITTER OPINION RELATED TO PERMENDIKBUD 30/2021 USING BIDIRECTIONAL LSTM

Zakiyatul Fitriyah[1], Mujiati Dwi Kartikasari[2*]

[1]*Department of Statistics, Universitas Islam Indonesia*
Kaliurang Street, Km 14.5 Sleman, Yogyakarta, 55584, Indonesia

*Corresponding author's e-mail: *mujiatikartikasari@uii.ac.id*

Article History:

Received: 24th February 2023
Revised: 9th May 2023
Accepted: 11th May 2023

Keywords:

Permendikbud 30/2021;
Classification;
LSTM;
BiLSTM.

ABSTRACT

During the COVID-19 outbreak, sexual violence in Indonesia has risen. Sexual abuse is prevalent even within the realm of education. Many incidents of sexual assault are reported within the higher education sector. The Ministry of Education, Culture, Research, and Technology published Decree of the Minister of Education and Culture Number 30/2021 (Permendikbud 30/2021) on the Prevention and Handling of Sexual Violence in Higher Education in an effort to prevent sexual violence on campus. This regulation's issuance has become a popular topic of discussion on social media. Twitter is one of the social media platforms where opinions are expressed. The publication of Permendikbud 30/2021 elicited a variety of views, from those who supported the rule to those who did not. This study's objective is to categorize tweets about Permendikbud 30/2021. Bidirectional LSTM (BiLSTM) was utilized to classify data in this study. The accuracy values are 87%, the precision values are 82%, and the recall values are 96% based on the findings of the analysis comparing training data of 80% to testing data of 20%.

How to cite this article:

Z. Fitriyah and M. D. Kartikasari., "TEXT CLASSIFICATION OF TWITTER OPINION RELATED TO PERMENDIKBUD 30/2021 USING BIDIRECTIONAL LSTM," *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 1113-1122, June, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

According to the Decree of the Minister of Education and Culture Number 30/2021 (Permendikbud 30/2021), sexual violence is any act of humiliating, harassing, and/or attacking a person's body and/or reproductive function, due to an imbalance of power and/or gender relations, which results in or can result in psychological suffering and/or interferes with a person's reproductive health and the opportunity to pursue higher education safely and optimally **[1].** Sexual assault is not limited to rape alone. In addition, every activity that leads to sexuality through coercion harms someone. Sexual assault can occur anywhere, including at home, in the surrounding environment, and even in the realm of education, which is a dangerous place to learn. Sexual violence can occur at all educational levels, from early childhood schooling through higher education. The highest number of reports of violence against women were received by the Komnas Perempuan from tertiary institutions, which ranked first **[2]**.

Not all instances of sexual violence on campuses are disclosed. From a survey conducted by the Ministry of Education and Culture in 2020 on lecturers, it was found that 77% of sexual violence occurred within universities, with 63% of them not reporting it **[3]**. For different reasons, some victims choose not to disclose or complain about sexual violence. The absence of rules that safeguard victims is one of them. Also contributing to campus confusion is the absence of rules and norms guiding the treatment of sexual violence on university campuses. Therefore, the Ministry of Education, Culture, Research, and Technology enacted Permendikbud 30/2021 on the Prevention and Handling of Sexual Violence in Higher Education in an effort to prevent sexual violence in universities.

The publication of Permendikbud 30/2021 on the Prevention and Handling of Sexual Violence has been a trending topic on Twitter. Twitter is a microblogging social media that Indonesians frequently utilize to communicate their thoughts via chirps, also known as tweets **[4]**. The community's responses to this vary. Multiple parties agree with and support the regulation to prevent sexual violence at school and foster a sense of safety on campus. However, many disagree with this rule and believe it could legalize adultery.

From tweets related to Permendikbud 30/2021, it can be classified into two categories, namely pros and cons. Classification is a process of find a model or function that explains or distinguishes concepts and data classes, with the aim of being able to estimate the class of an object whose class is unknown **[5]**. There are several classification methods. One method that can be used for classification is bidirectional LSTM (BiLSTM).

BiLSTM is a deep learning method that is an extension of long short-term memory (LSTM). BiLSTM has two layers whose processes are reversed, namely the lower layer that moves forward (forward) and understands and processes from the first word to the last word, and the layer above it that moves backward (backward), namely understanding and processing from the last word to the first word **[6]**. The advantage of this method is that it can understand the context of the sentence and represent the text better because it has a two-way layer that can take perspective from the previous word and the leading word.

This study aims to classify tweets that support (pro) and reject (contra) the regulation of the Minister of Education and Culture Number 30 of 2021 concerning the Prevention and Handling of Sexual Violence (PPKS) using the Bidirectional Long Short-Term Memory (BiLSTM) method.

2. RESEARCH METHODS

The data used in this study was obtained from the results of crawling on Twitter related to Permendikbud 30/2021 in Indonesia in November 2021. From the crawling results, 10,000 tweets were obtained. The data will be pre-processed before analysis and will then be categorized into two, namely pros and cons. The flowchart that is annexed to **Figure 1** shows the research methodology that was used to carry out this study.

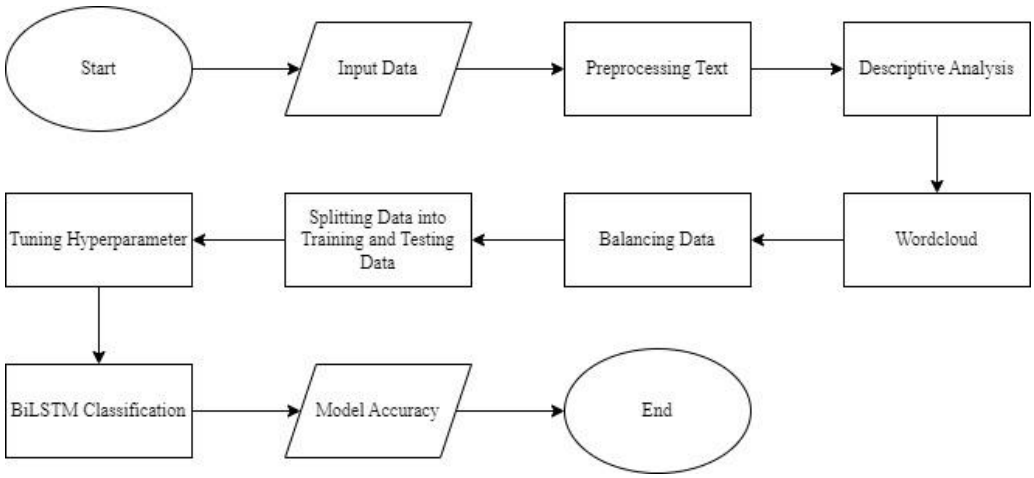

Figure 1. Research flowchart

2.1. Text Mining

Text mining is an important data mining technique that includes the best ways to find patterns [7]. Text mining is a method of conducting knowledge searches that focuses on data in the form of documents or texts with the goal of extracting and identifying useful information. Raw data from the results of text mining activities must be prepared in advance according to the needs of the analysis so that it can produce good analytical results [8]. This is where initial processing or preprocessing of text data is needed.

2.2. Deep Learning

Deep learning can be considered the development of an artificial neural network. In deep learning, a system can learn to classify directly from images or voice [9]. Deep learning is very well suited to apply to supervised learning, unsupervised learning, and semi-supervised learning, as well as for reinforcement learning in various applications like the introduction of the image of, the sound of, classifications, text, and so on. Modeled on the DL, it is basically built on an artificial neural network (neural network). If a network has more than three layers, the network includes a deep network [10].

2.3. Synthetic Minority Oversampling Technique (SMOTE)

The synthetic minority oversampling technique (SMOTE) is one method that can overcome data imbalances. This method increases the number of datasets in the minor class by generating synthesis data so that the number of datasets for the minor class and major class are balanced. Synthetic data is created based on k-nearest neighbors. Synthesis data generation for numeric and categorical data is different. Numerical data is measured by its proximity to the Euclidean distance, whereas categorical data is measured by the mode value. The calculation of the distance between examples of minor classes whose variables are on a categorical scale is carried out using the Value Difference Metric (VDM) [11].

2.4. Bidirectional Long Short-Term Memory (BiLSTM)

Long Short-Term Memory (LSTM) is the development of the recurrent neural network (RNN) architecture to deal with the vanishing gradient problem [12], where the slope of the loss function decreases exponentially when processing long sequential data. This problem causes the RNN to fail to capture long-term dependencies so that it can reduce prediction performance [13]. LSTM replaces the RNN layer with a memory cell block using a gate mechanism consisting of a forget gate, an input gate, and an output gate. Similar to an RNN, a LSTM is composed of neurons that are processed repeatedly. Figure 2 is the structure of a single neuron in LSTM [12].

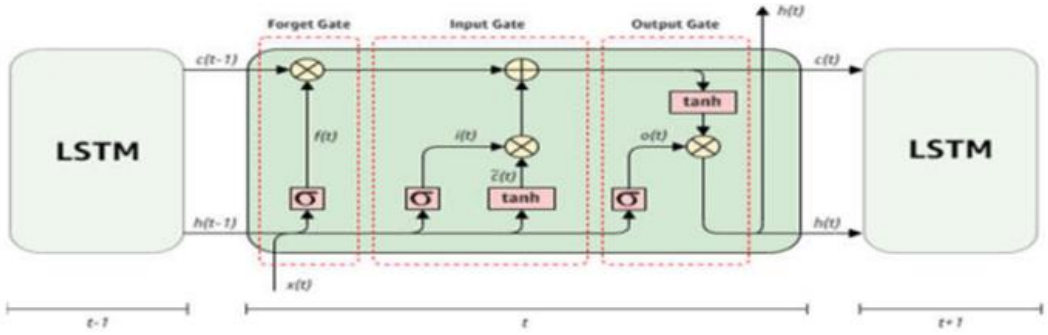

Figure 2. Single neuron on LSTM architecture

The forget gate is the first gate on the LSTM that accepts input h_{t-1} and x_t to produce a value of 0 or 1 in C_{t-1}. When the forget gate is 1, the cell state will store information, whereas if the value is 0, the information will be removed from the cell state. The input gate is the second gate in the LSTM to determine what information will be stored in the cell state consisting of the sigmoid layer and the tanh layer. The sigmoid layer decides which value to update, and the tanh layer creates a new value C_t to be added to the cell state. The output gate is the last gate on the LSTM and determines the output of the cell state [11]. The LSTM flow is arranged in the following equation [14].

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

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

$$\widetilde{C_t} = \tanh\,(W_c[h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = ft * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t * \tanh\,(C_t) \quad (6)$$

where W_f, W_i, W_c, W_o is trained weights, b_f, b_i, b_c, b_o is trained biases, σ is function sigmoid, x_t is input in time t, h_{t-1} is hidden state previous time, C_t is cell state in time t, h_t is output in time t, f_t is forget gate in time t, i_t is input gate in time t, o_t is output gate in time t.

LSTM only reads sentences in one direction, from beginning to end [15]. Bidirectional long short-term memory (BiLSTM) is a development of the LSTM model where there are two layers whose processes are in opposite directions to each other. This model is very good for recognizing patterns in sentences because each word in a document is processed sequentially, because reviews can be understood if learning is sequential for every word. The layer below moves forward, that is, understands and processes from the first word to the last word, while the layer above moves backward, that is, understands and processes from the last word to the first word [6]. The two hidden states $h_t^{forward}$ and $h_t^{backward}$ from the LSTM are combined to form the final hidden state h_t^{BiLSTM}, as expressed below

$$h_t^{BiLSTM} = h_t^{forward} \oplus h_t^{backward}. \quad (7)$$

With this layer of two opposite directions, the model can understand and take perspective from the previous word, and the leading word, so that the learning process will be deeper, which will impact the model's ability to better understand the context in the review. **Figure 3** is BiLSTM Architecture.

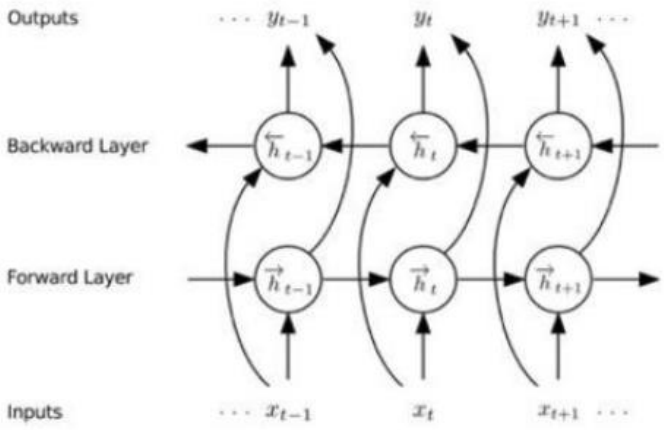

Figure 3. BiLSTM architecture

3. RESULTS AND DISCUSSION

3.1. Preprocessing Data

The dataset in this study was obtained from the results of crawling social media related to Permendikbud 30/2021 in Indonesia in November 2021. 10,000 tweets were obtained from the crawling results. The data will be preprocessed before analysis. The data obtained from the results of the crawling process still contains capital letters, characters or punctuation marks, incorrect spelling, and words that are less meaningful, which can affect the analysis. Thus, a preprocessing process is needed to clean the data. There are several preprocessing stages. The first is cleaning, namely removing symbols and punctuation marks, numbers, URLs, hashtags, usernames, and RTs from tweet data. Case folding, which is changing the use of capital letters to lowercase letters or lower case on tweet data; and filtering, which is removing words that are less needed and do not convey a significant message such as pronouns, conjunctions, and others in a text based on a stopword dictionary. It is normalizing, namely normalizing non-standard words or words in foreign languages into words that are in accordance with the KBBI. The next stage is the tokenizing stage, which involves breaking sentences into words that do not affect each other, which are called tokens, and finally the stemming stage is the stage to change words to their basic form by removing affixes.

The data from the crawling results still contain duplicate or twin data caused by tweets that are posted repeatedly. So, it is necessary to delete duplicate data manually. After deleting duplicate records, 1926 records were obtained from 10,000 records. Furthermore, the Permendikbud data were manually labeled into two categories, namely "pro" and "con." This labeling uses the researcher's point of view. A tweet or tweet is included in the "pro" category if the tweet contains positive words, not badmouthing, and words of support for Permendikbud 30/2021. Meanwhile, a tweet or tweet falls into the "contra" category if the tweet contains negative words and tends to vilify Permendikbud 30/2021. Based on the labeling results, 1609 data points are in the pro category, and 317 are in the con category. **Table 1** is the data that has been labeled.

Table 1. Twitter opinion related to Permendikbud 30/2021

No.	Tweet	Category
1.	arah kampus terjadi pengajian kampus mentoring islam halangi dosen berprestasi disingkirkan isu radikalradikul pelajaran agama islam dua semester jam perminggu sementara sekulerisasi liberalisasi seksual legal permendikbud	con
2.	permendikbud terbukti efektif para korban selama bungkam akhirnya berani suara artinya masyarakat anggap permendikbud layak tumpuan harapan hapus kekerasan seksual	pro

The datasets that have gone through text preprocessing and labeling are then separated by spaces using a tokenizer from the Keras library. Next, the list of vocabulary tokens is converted to a numerical sequence by replacing the index of each vocabulary with an integer value.

3.2. Balancing Data

The research data held are unbalanced data because the amount of data in the class is less or more than the number of other classes, which results in inequality in one class. In classification, imbalanced data can affect the performance of the analysis. One method to handle imbalanced data is SMOTE. The SMOTE method is to balance the amount of data by increasing the number of minority class data by generating random data so that it is balanced with the number of majority class data. **Figure 4** is a data plot before data balancing with SMOTE is carried out.

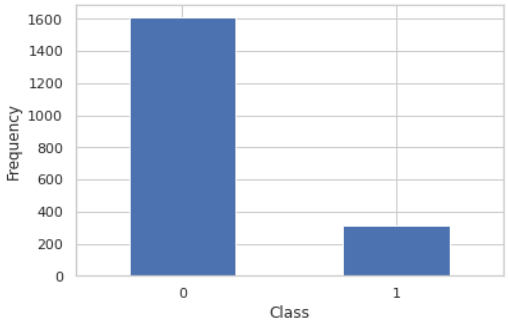

Figure 4. Data plot before data balancing using SMOTE

Figure 4 shows a plot prior to SMOTE, with the number of data points labeled "pro" at 1609 and the number of data points labeled "con" at 317. It can be seen that the difference in the amount of data between the pro and con categories is quite significant. So, to avoid errors in classification, data balancing is carried out with SMOTE so that the amount of data is balanced. **Figure 5** is the plot after SMOTE.

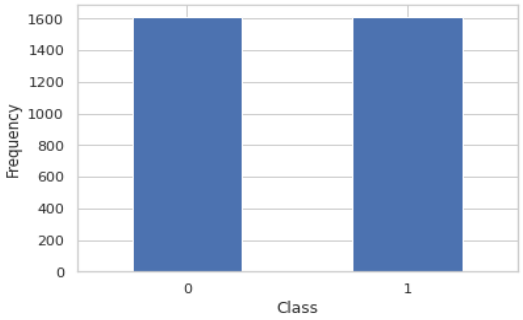

Figure 5. Data plot after data balancing using SMOTE

Figure 5 shows the plot after SMOTE. After balancing the data using SMOTE, the amount of data in the "con" category with label 1 increased from 317 to 1609 label 1 increased from 317 to 1609. It can be seen that the "con" category, which is the minority category, has balanced the amount of data on the majority class, namely "pro."

3.3. Splitting Dataset

After the data balancing process has been carried out, the next process is splitting the dataset into training data and testing data. Training data aims to train the algorithm. Testing data is data used to test the accuracy and performance of the previously trained model. In this study, we used a comparison of 80% for training data and 20% for test data.

The total data after data balancing is 3218. After sharing 80% of the training data, a total of 2574 training data sets are obtained, with 1287 in each category. Then, for 20% of the test data obtained, 644 were obtained, with 322 data for each category ("pros") and "cons" obtained.

3.4. Modelling Data Using Bidirectional LSTM

The next process is the classification of tweets that are included in the "pro" and "con" categories to get the best model using the bidirectional LSTM method. Before classification with bidirectional LSTM, it is necessary to first make a model with several parameters. The model made is a sequential type, which is a model made on a layer-by-layer basis. In addition, the researchers did hyperparameter tuning to obtain the

best parameters from the bidirectional LSTM model in order to produce better performance. The GridSearchCV tuning method was used in this study. The best hyperparameter tuning parameters will be used to create a new model that will outperform the previous model in classification. **Table 2** are the hyperparameters used in the tuning and the optimal parameters obtained.

Table 2. Tuning hyperparameter

No.	Tuning hyperparameter	Parameter optimal
1.	embed_dim: [32, 64]	embed_dim: 64
2.	hidden_unit: [16, 32, 64]	hidden_unit: 16
3.	dropout_rate: [0.2]	dropout_rate: 0.2
4.	optimizers: [Adam, RMSprop]	optimizers: Adam
5.	learning_rate: [0.01, 0.001, 0.0001]	learning_rate: 0.001
6.	epochs: [10,25]	epochs: 25
7.	batch_size: [128, 256]	batch_size: 128

The classification process is done on the training data using the BiLSTM method with optimal parameters found through hyperparameter tuning. This is done to train the algorithm and study data patterns from the training data. **Figure 6** is a plot of how well the classification worked when BiLSTM was used with training data.

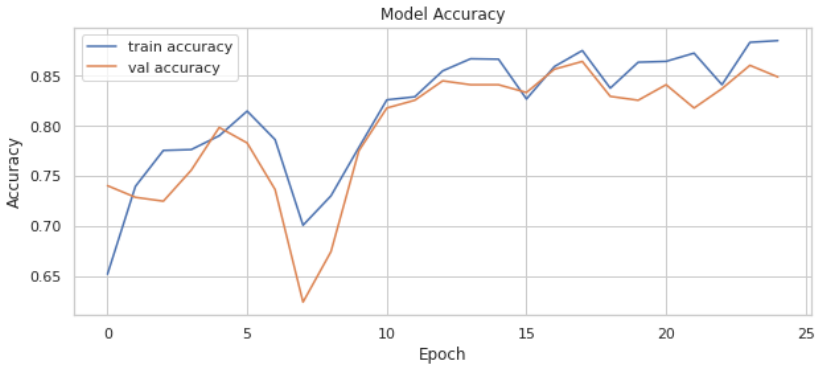

Figure 6. Plot the results of classification accuracy with training data

Based on **Figure 6**, it can be seen that the training accuracy increases and decreases with each increase in epoch. The highest accuracy is achieved when the number of epochs is 25. For the accuracy of the training data, the classification results have an accuracy of 88.51%.

3.5. Evaluation Model

After classifying the training data using optimal parameters, the classification process is carried out on the test data to see the performance and accuracy of the previously trained algorithm. the BiLSTM method. **Table 3** is the result of the confusion matrix using the BiLSTM method.

Table 3. Confusion matrix

Actual	Prediction	
	Pro	Con
Pro	306	12
Con	69	257

There are 306 correct data points in the pro category and 257 correct data points in the con category. From the results of the confusion matrix, calculations were made to obtain the values of accuracy, precision, and recall.

a) Accuracy

Accuracy is a value that indicates the level of closeness between the predicted value and the actual value.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$= \frac{306 + 257}{(306 + 257 + 12 + 69)}$$

$$= 0.87$$

b) Precision

Precision is the ratio of the prediction of a selected relevant item to the entire selected item

$$precision = \frac{TP}{TP + FP}$$
$$= \frac{306}{306 + 69}$$
$$= 0.82$$

c) Recall

Recall is the prediction ratio of a selected relevant item to the total number of items.

$$recall = \frac{TP}{TP + FN}$$
$$= \frac{306}{306 + 12}$$
$$= 0.96$$

4. CONCLUSIONS

Based on the results and discussions that have been explained, we have some conclusions.

a). The data used is tweet data regarding Permendikbud 30/2021, which was taken from Twitter using a crawling technique. We obtained 10,000 records from the crawling results. After manually preprocessing and removing duplicate data, 1926 records were obtained. The information was divided into two categories: pros and cons. The data for the pro category is 1609, and the data for the contra category is 317.

b). Classification analysis employs the bidirectional LSTM method, which compares 80% of training and 20% of test data, respectively. To get optimal parameters, hyperparameter tuning is used. The optimal parameters for classification are batch size: 128, dropout: 0.2, embedding: 64, epochs: 25, hidden unit: 16, learning rate: 0.01, and optimizers: Adam. These parameters yielded 87% accuracy, 82% precision, and 96% recall.

REFERENCES

[1] Kemendikbudristek, “Abstrak Permen 30 Tahun 2021 Tentang Pencegahan Dan Penanganan Kekerasan Seksual Di Lingkungan Perguruan Tinggi,” vol. 5, p. 6, 2021.

[2] Komnas Perempuan, “Kekerasan Seksual di Lingkungan Pendidikan,” 2020.

[3] Kemendikbud, “Wujudkan Lingkungan Perguruan Tinggi yang Aman dari Kekerasan Seksual,” 2021. .

[4] A. Novantirani, M. K. Sabariah, and V. Effendy, “Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine,” *e-Proceeding Eng.*, vol. 2, 2015.

[5] I. Budiman and R. Ramadina, “Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi,” *Ijccs*, vol. x, No.x, no. 1, pp. 1–5, 2015.

[6] D. J. M. Pasaribu, K. Kusrini, and S. Sudarmawan, “Peningkatan Akurasi Klasifikasi Sentimen Ulasan Makanan Amazon dengan Bidirectional LSTM dan Bert Embedding,” *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 9–20, 2020, doi: 10.35585/inspir.v10i1.2568.

[7] E. M. O. N. Haryanto, A. K. A. Estetikha, and R. A. Setiawan, “Implementasi SMOTE untuk Mengatasi Imbalanced Data Pada Sentimen Analisis Sentimen Hotel di Nusa Tenggara Barat dengan Menggunakan Algoritma SVM,” *J. Inf. Interaktif*, vol. 7, 2022.

[8] F. F. Rachman and S. Pramana, “Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter,” *Heal. Inf. Manag. J. ISSN*, vol. 8, no. 2, pp. 2655–9129, 2020.

[9] H. A. Parhusip, “Study on COVID-19 in theWorld and Indonesia Using Regression Model of SVM, Bayesian Ridge and Gaussian,” *J. Ilm. Sci.*, vol. 20, 2020, doi: 10.35799/jis.20.2.2020.28256.

[10] I. Cholissodin and A. A. Soebroto, “AI , Machine Learning & Deep Learning (Teori & Implementasi),” no. December, 2021.

[11] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, “SMOTE: Potensi Dan Kekurangannya Pada Survei,” *E-Jurnal Mat.*, vol. 10, no. 4, 2021.

[12] K. S. Nugroho, I. Akbar, A. N. Suksmawati, and I. Istiadi, “Deteksi Depresi dan Kecemasan Pengguna Twitter,” *4th Conf. Innov. Appl. Sci. Technol. (CIASTECH 2021)*, no. Ciastech, pp. 287–296, 2021.

[13] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, “LSTM network: a deep learning approach for short-term traffic forecast,” *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.

[14] J. Abdillah, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Emosi pada Lirik Lagu menggunakan Metode Bidirectional LSTM dengan Pembobotan GloVe Word Representation," *J. Resti*, vol. 5, no. 1, pp. 1–4, 2020.
[15] H. Elfaik and E. H. Nfaoui, "Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text," *J. Intell. Syst.*, vol. 30, no. 1, pp. 95–412, 2021.