Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 8, No. 1, March 2022, pp. 64-75
ISSN: 2338-3070, DOI: 10.26555/jiteki.v8i1.23009

Sentiment Analysis and Topic Modelling of The COVID-19 Vaccine in Indonesia on Twitter Social Media Using Word Embedding

Kartikasari Kusuma Agustiningsih, Ema Utami, Omar Muhammad Altoumi Alsyaibani
Magister of Informatics Engineering, Universitas Amikom Yogyakarta

ARTICLE INFO

Article history:

Received January 17, 2022
Revised February 17, 2022
Accepted March 01, 2022

Keywords:

Sentiment Analysis;
Topic Modeling;
Word Embedding;
Fasttext;
GloVe;
Latent Dirichlet Allocation

ABSTRACT

This study aims to analyze the sentiments of the Indonesian people towards the COVID-19 vaccine on Twitter. Data collection was carried out from September 2020 to June 2021 with the keyword "covid vaccine," which resulted in 262306 tweets. After filtering and cleaning, there are 83384 tweets left. The labeling process was done manually by an expert. The label composition in the data is 35209 tweets of positive sentiment, 41596 tweets of neutral sentiment, and 6579 tweets of negative sentiment. The remaining data is preprocessed using case folding, removing punctuation, stopword removal, stemming, and the application of slang words. The highest number of tweets appeared in January 2021, after Joko Widodo became the first person in Indonesia to receive a vaccine injection. The number of tweets reached 23492 tweets. At the topic modeling stage, measurements were conducted using the Coherence Score. The distribution of the optimal number of topics is 3 topics. The first topic, with a token percentage value of 51.8%, leads to positive sentiment, while the second and third topics, with token percentage values of 24.5% and 23.7%, lead to neutral sentiment. Bidirectional LSTM architecture was implemented to perform sentiment classification. Fasttext and GloVe word embedding was tested to vectorize tweet data. The test accuracy generated by Fasttext word embedding reached 75,7690%, while the test accuracy produced with GloVe word embedding reached 74.7017%. The usage of slang words could not increase the test accuracy in this study. The use of the Modelcheckpoint to monitor model performance during training could produce a model with a slightly higher test accuracy, about 1.07% (in scenario 1 and scenario 6), compared to a model whose performance was monitored using Early Stopping. In future research, it can be tried to apply a lower learning rate to produce better accuracy in a large number of epochs, or it could be by changing the dropout parameter.

Corresponding Author:

Kartikasari Kusuma Agustiningsih, Magister of Informatics Engineering, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
Email: kartikasarikusuma@students.amikom.ac.id

1. INTRODUCTION

Corona Virus Disease began to spread in Wuhan at the end of 2019 [1]. This virus is later known as COVID-19. COVID-19 can spread through the air and pass from one person to another [2]. COVID-19 entered Indonesia in February 2020 and was only determined by the Government of Indonesia in early March 2021 [3]. Since then, COVID-19 has spread to all provinces in Indonesia. Vaccines are products or substances that are inserted into the body by injection or by mouth to stimulate the body's immune system in dealing with certain diseases [4]. The COVID-19 vaccine began to be developed in early 2020. Author [5] stated that 115 vaccine candidates had been proposed in April 2020. The Government of the Republic of Indonesia has successfully collaborated with Sinovac, Sinopharm, G42 Health Care, CanSino, Genexine, and COVAX in

efforts to meet the needs of the COVID-19 vaccine for the Indonesian people [6]. In addition, the government was also developing a domestic vaccine under the "Merah Putih" banner [6].

On the other hand, many hoaxes and negative news circulated about vaccines, both about the safety and halalness of vaccines [7]. The author [7] asserted that conservative handling of rumors and hoaxes has no impact on millennials and digital natives who constantly consume information from social media. Therefore, it is necessary to conduct research on public opinion sourced from social media. Twitter is a text-based social media with up to 140 characters per post [8]. Twitter has been widely used as a source of data mining, especially in the fields of Natural Language Processing (NLP) and Sentiment Analysis research [9].

Several studies have been done regarding the sentiment analysis of Indonesian people towards the COVID-19 vaccine, and data are sourced from Twitter. For instance, the study by authors [10] conducted sentiment analysis on two types of vaccines, namely Sinovac and Pfizer. The study collected data from Twitter. The classification methods used were Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). This study did not mention the vectorization techniques used. The highest accuracy was performed by SVM, which reached 85%.

Another study that also used the Naive Bayes method is research [11], in which 3780 tweets were collected and classified into 3 classes, namely positive, negative and neutral. Term Frequency-Inverse Document Frequency (TF-IDF) weighting was used in the study to score each word that appeared. The results of this study stated that positive sentiment is greater than negative sentiment. The naive Bayes method resulted in 93% of accuracy. The TF-IDF weighting was also used in the study [12] to perform word weighting on 488 data sourced from Twitter. Data were collected from 13 to 20 January 2021. The study performed a binary classification on sentiment analysis of the COVID-19 vaccine using the Naive Bayes method. Public sentiment towards the COVID-19 vaccine almost balances 51.4% of data containing positive sentiment and 48.6% of data containing negative sentiment. The study suggests that the future research preprocessed data using a slang word dictionary because Twitter users often do not use formal language.

Differently, research [13] compares the Recurrent Neural Network (RNN) and Naive Bayes in classifying the sentiment analysis of the COVID-19 vaccine on a dataset collected from Twitter from January to April 2021. The data collection resulted in 5000 tweet data. From the comparison results, the authors found that the RNN method used performed with higher accuracy than Naive Bayes. Several previous studies that conducted sentiment analysis toward the COVID-19 vaccine on the Indonesian corpus mostly used TF-IDF weighting to vectorize words. Therefore, this study aims to implement word embedding as a vectorization technique. Research [14] stated that word embedding could improve the performance of traditional vectorization techniques such as a bag of words. Several previous studies have studied sentiment analysis on the Indonesian corpus using a word embedding to vectorize words. For example, research [15] used a probabilistic neural network method to conduct sentiment analysis on responses to the Sinovac vaccine on social media Twitter. The study used Fasttext word embedding Fasttext to vectorize data. The study found that the word embedding used can improve model accuracy, which reached 85%.

The research conducted by authors [16] classified the public sentiment towards the COVID-19 vaccine using the RNN method on a dataset sourced from social media Twitter. The data were classified into two classes, namely positive and negative. The study compared the performance of Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM. In the study, the highest accuracy was obtained by the Bidirectional LSTM method, which reached 91%. Unfortunately, this study did not describe the vectorization technique used.

In this study, Fasttext and GloVe word embedding would be used for data vectorization. In addition, according to the results of the study [16] about Bidirectional LSTM method performance, the method was also implemented in this study to classify sentiment data. Therefore, the contribution of this research is to examine public sentiment every month in the form of a timeline that has never been done by previous studies and capture sentiment data from the beginning of the discussion about vaccines on Twitter until June 2021, which is when the Indonesian government disseminates vaccines to the public. By capturing tweet data over a long period of time, this research will also contribute to gathering a large amount of vaccine sentiment data from Twitter.

2. METHOD

2.1. Data Collection

The data collection technique used in this study was the SNScrape tool which was introduced by [17]. SNScrape requires users to have a Twitter developer account to be able to scrape data. This tool can scrape data more than 7 days in advance. The scraped data were data from 1st September 2020 to 31st June 2021 using the keyword "vaksin covid." Data collection was conducted in stages because the data obtained was

plentiful. The data which has been collected were labeled manually. The data collection process can be seen in Fig. 1.

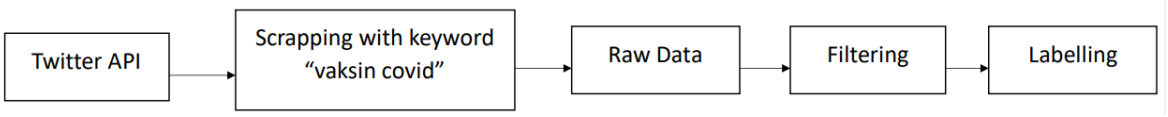

Fig. 1. Data Collection Process

2.2. Data Preprocessing

The collected data still needs to be preprocessed because it was still raw data. The first step was filtering the data. Data filtering was done by removing duplicate data, removing tweet data from news accounts and organizations, and removing tweet data in the form of questions because they do not represent public sentiment. The filtered tweet data were manually labeled. Tweet data that are not in the Indonesian language found at the labeling stage were deleted.

The next preprocessing step was converting all tweet data into lowercase letters (case folding). According to the author [18], the most widely used preprocessing techniques in sentiment analysis of the COVID-19 vaccine are stopword removal and removing punctuation. Stopword removal is a technique to remove irrelevant or potentially irrelevant words [10]. Meanwhile, removing punctuation removes characters, tags, mentions, and URLs. These two techniques were also applied in this study.

Stemming is a technique that returns words to basic words. The stemming technique used in this study was the stemming algorithm introduced by [19]. Some Twitter users in Indonesia use non-standard language (slang) in making tweets [20]. In this study, a slang word dictionary published by authors [21] was tested. The generated data at this stage would be classified and used in the topic modeling process. The workflow of this research is shown in Fig. 2.

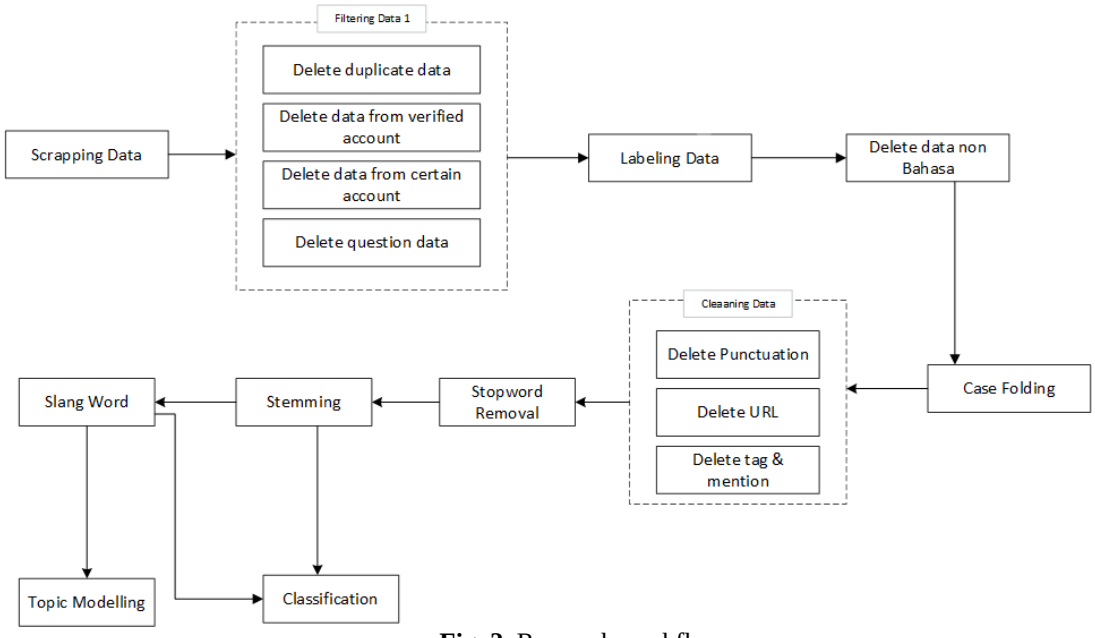

Fig. 2. Research workflow

2.3. Word Embedding

In order for a model to process a sentence in a natural language processing study, a word must be converted into a vector form. GloVe and fastText are techniques that can convert words into vectors. The gloVe is an algorithm introduced by [22]. The gloVe is based on the factorization technique [23]. It is referred to as a modification of Word2vec [24]. The GloVe method is formulated by the authors [22] into Equation 1.

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij}\right)^2 \quad (1)$$

Where V represents corpus size, b represents bias, w represents weight, and X represents word-processed on matrix $i \times j$.

FastText is a Facebook-owned library used to generate efficient word representations and provide support for text classification [25]. fastText is generally used to solve sentence classification and word representation problems to be more efficient and faster than the Word2vec and GloVe methods [26]. In fastText, this word vector is then stored in two files, the .bin file and the .vec file [27]. fastText function is represented in Equation 2. S represents a scoring function, w represents weight, l represents $log(1 + e - x)$, and n represents the number of words in the corpus.

$$\sum_{t=1}^{T}\left[\sum_{c\in Ct} l(s(w_t, w_c)) + \sum_{n\in N_{t,c}} l(-s(w_{t,}n))\right] \quad (2)$$

2.4. Topic Modeling and Classification Methods

Topic modeling is a technique that can identify topics from a set of documents containing a collection of words. The topic modeling model used in this study was Latent Dirichlet Allocation (LDA), published by [28]. To detect latent topics in large textual data, LDA is very useful [29]. In the LDA concept, documents are presented by topics where each topic is characterized by a word distribution. In order to determine the number of topics that should be explored from a corpus, a Coherence Score calculation was implemented. This method measures the subject score by measuring the degree of semantic similarity between high-scoring words in the subject. This measurement helps the model distinguish between subjects that can be interpreted semantically and those that have numerical inference artifacts.

In order to test the performance of word embeddings, the Bidirectional LSTM network was created for each test scenario. The dataset obtained was divided into 2 parts, training validation data, and test data with a distribution of 80:20. The models were created using the TensorFlow and Keras frameworks. The performance of the model during training was monitored using Early Stopping with a patience value of 3 and a Model checkpoint. The Early Stopping and Modelcheckpoint monitored validation loss during the training process. If the Early Stopping found that validation loss increased for 3 iterations during training, then the training model would be stopped. The Modelcheckpoint callback is flexible in the way it can be used, but in this case, we will use it only to save the best model observed during training as defined by a chosen performance measure on the validation dataset.

Word embeddings implemented used 300 dimensions. In all scenarios, the same Bidirectional LSTM architecture was implemented. First of all, the data will enter into the Input layer and then be forwarded to the Embedding layer. In the next stage, 3 Bidirectional LSTM layers were implemented with 128, 64, and 32 neurons, respectively. In the last layer, the Dense layer was deployed containing 3 neurons with Softmax activation function because the dataset had 3 classes. To avoid the model from being overfitted, the Dropout layer was implemented in between each layer. The weight of each neuron was optimized during the training process using Adam developed by the authors [30]. According to the author [31], Adam's optimization function can make the model achieve the optimal weight value in a short number of training iterations. During the training process, 30% of the training data was used as validation data. The batch size value used was 64. The preprocessing, word embedding, topic modeling, classification, and training monitoring methods used in this study are illustrated in Fig. 3.

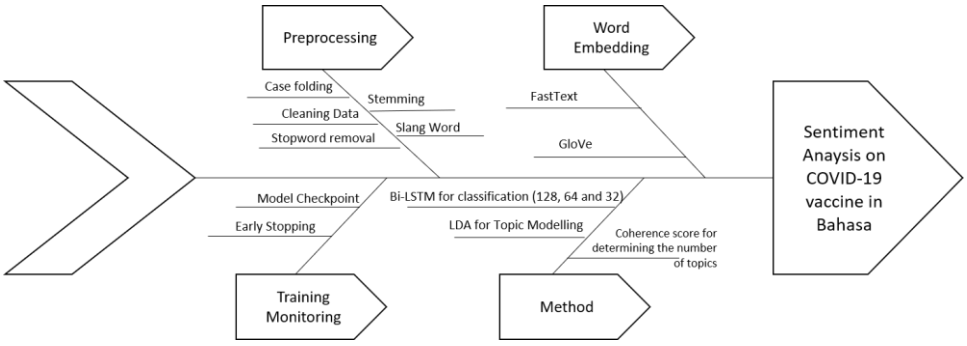

Fig. 3. Research Methods

3. RESULTS AND DISCUSSION

3.1. Dataset

Data collection was carried out from September 2020 to June 2021 using the SNScrape tool. The data collected amounted to 262306 tweets. The data obtained identified as many as 872 organizational accounts and news sites. In this study, data originating from these accounts was omitted with the aim that the data processed only came from ordinary people.

Data collection was conducted in several stages due to the limitations of the Twitter API. This has an impact on the possibility of duplicate tweets. The identification of duplicate tweets was based on the tweet_id feature from the scraping results. Only one tweet datum was kept from each duplicate tweet found. The remaining tweet data from this process became 91151 tweets. The data were labeled manually by the expert. At the time of labeling, tweet data that were not in Indonesian were still found, so the data needed to be removed and resulting in 83384 tweet data.

The remaining data were then preprocessed. Preprocessing steps consist of case folding, cleaning data (remove punctuation, remove URL, remove mention and tag), stopword removal, and stemming. Finally, the data were preprocessed using a slang words dictionary published by the author [21]. The data later would be used in classification and topic modeling.

3.2. Sentiment Analysis

From the filtering results, there are 83384 tweets consisting of 35209 tweets with positive sentiment, 41596 tweets with the neutral sentiment, and 6579 tweets with negative sentiment. This can be seen in Fig. 4, where positive sentiment data is represented in green, negative sentiment data is represented in red, and neutral sentiment data is represented in blue. From this data, it can be observed that 50% of tweets are neutral. On the other hand, only 8% of people's tweets have negative sentiments, while those with positive sentiments are 42%. From the figure, it can be concluded that the rejection of vaccines in Indonesia by people who use Twitter was still relatively low. However, public discussion about vaccines was quite high, where neutral data reached half the existing data. Furthermore, the number of people who have positive sentiments about vaccines is more than five times the number of people with negative sentiments.

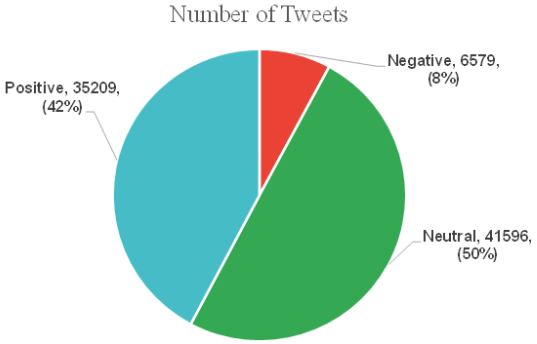

Fig. 4. Number of tweets collected based on sentiment

In order to understand the data further, the data were grouped by month. This data is represented in Fig. 5. From this data, it can be identified that public tweets about vaccines since the beginning of the discussion about “vaccines” by Indonesian people on Twitter were below 4000 tweets. The intensity of this discussion began to increase in December 2020, which reached 7896 tweets when the Minister of Health of the Republic of Indonesia was changed [32], and there was an increase in the number of active cases in Indonesia [33]. On the other hand, in the same month, Joe Biden won the United States Presidential election and received his first vaccine injection [34]. The number of tweets about vaccines reached the highest point in January 2021 when the President of the Republic of Indonesia became the first person in Indonesia to be injected with covid vaccine [35]. The number of tweets reached 23492 tweets.

In February 2021, public discussion of vaccines on Twitter decreased drastically compared to the previous month but was still quite high when compared to the first three months when vaccine discussions started on Twitter. This was because, in that month, the Vaksin Nusantara was being researched and developed. On the other hand, the first three quarters of 2021 would be the implementation of phase 1 vaccination specifically for health workers. This discussion began to increase again in March until it reached 11608 tweets when the public began to be registered to get the first dose of the vaccine [36]. The discussion spiked again in June 2021, when there was a significant increase in active cases [33] and the spread of the Delta variant of covid [37].

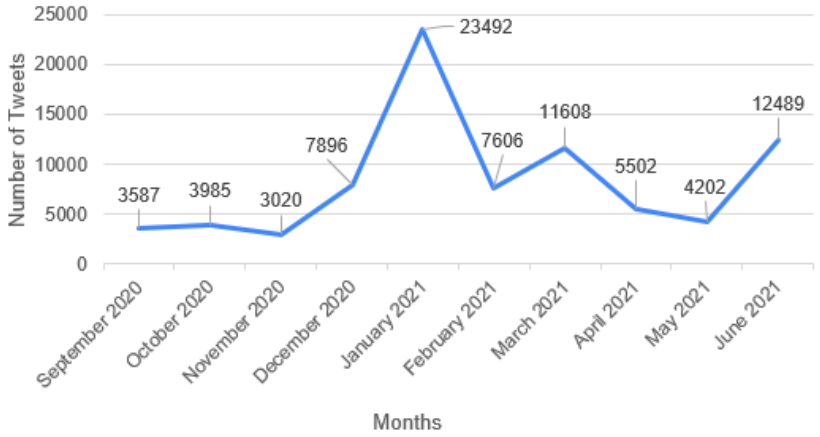

Fig. 5. Tweet data per month

Table 1 and Fig. 6 explain monthly public sentiment data from September 2020 to June 2021. Based on Table 1 and Fig. 6, it can be observed that the number of tweets with negative sentiments is always lower than the number of tweets with positive or neutral sentiments each month. Moreover, the number of tweets with positive and neutral sentiments is not much different every month. The number of tweets with neutral sentiment is always more than the number of tweets with positive sentiment except in November 2020, May 2021, and June 2021. In November 2020, PT. Bio Farma has received the Bulk Production of the COVID-19 vaccine, which was the result of an agreement between the Government of the Republic of Indonesia and the Sinovac vaccine manufacturer [38]. Meanwhile, from April 2021 until June 2021 was the second stage of vaccination for the public [36]. From this data, it can be concluded that although there are Indonesian people who respond negatively to vaccination, the number is very low when compared to people who respond to vaccination activities with positive sentiments.

Table 1. Tweet data grouped by month

Month	Number of Positive Tweets	Number of Negative Tweets	Number of Neutral Tweets
September 2020	1143	429	2012
October 2020	1449	616	1920
November 2020	1518	523	979
December 2020	3273	634	3989
January 2021	11406	1879	10207
February 2021	2987	403	4216
March 2021	5436	479	5693
April 2021	2612	382	2508
May 2021	1185	410	2608
June 2021	4200	824	7465

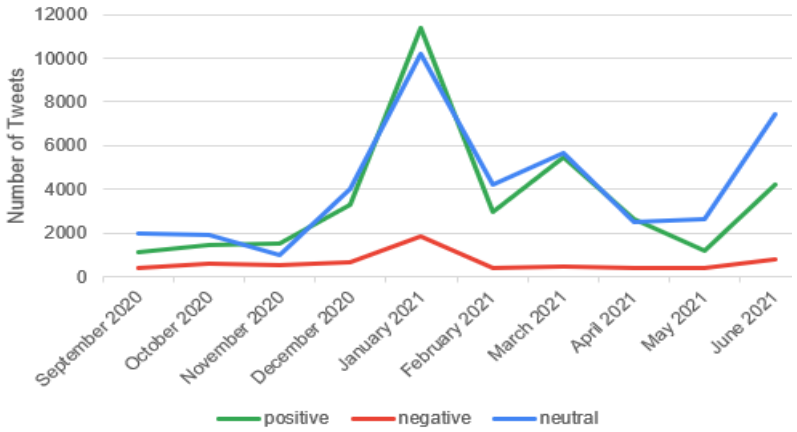

Fig. 6. Tweet data grouped by month

3.3. Topic Modelling

The Coherence Score was measured in the corpus before determining the number of topics to be explored in the topic modeling phase. Measurement was limited to 10 topics and trained in 300 iterations. The measurement results can be seen in Fig. 7. Based on Fig. 7, it can be seen that the highest coherence value was obtained by dividing the corpus into 3 major topics. The coherence value reached 0.49 in the 3 topics distribution.

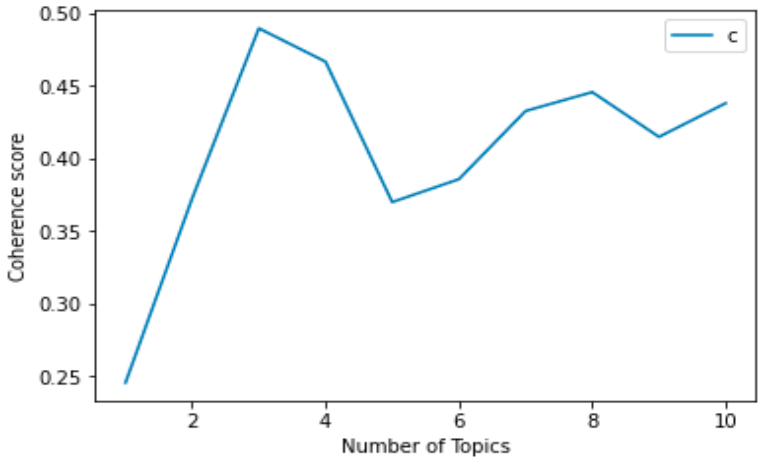

Fig. 7. Coherence Score on Topic Modelling

Table 2 shows the most relevant words for each topic. Based on Table 2, it can be understood that the first topic discusses the opinion of people who want vaccines and regret that there were other people who do not want vaccines because they are afraid of getting sick. In addition, this topic also discusses the possibility of being immune to the virus after being vaccinated. The second topic is about calls to participate in mass vaccination activities carried out by the TNI and POLRI in several provinces. In addition, there is also a discussion about whether vaccination at the time would invalidate the fast or not. The third topic is a discussion about efforts to break the chain of covid spread by various parties by way of disciplined health protocols and PPKM (Enactment of Community Activity Restrictions).

Table 2. Top-30 relevant words in each topic

Topic	Top-30 relevant words	Token Percentage
Topic 1	yang, enggak, sudah, kena, mau, orang, saja, kalo, gue, percaya, jadi, aku, banget, apa, bisa, kalau, karena, kayak, bukan, tapi, nya, sekarang, kebal, deh, sih, habis, positif, malah, buat, kan	51.8%
Topic 2	vaksinasi, massal, dosis, juta, rangkai, nasional, indonesia, sinovac, guna, covid, ayo, polri, vaksin, lawan, aman, perintah, astrazeneca, hari, halal, program, pandemi, dukung, tahap, masyarakat, tiba, puasa, produksi, juni, sebut, lansia	24.5%
Topic 3	bhayangkara, giat, sebar, desa, jaga, cegah, patuh, prokes, puskesmas, dokter, protokol, polda, koramil, tempat, polsek, babinsa, terap, camat, damping, hindar, bhabinkamtibmas, kab, waspada, haji, kec, ppkm, vaks, laksana, mata, sehat	23.7%

The first topic is quite dominant compared to other topics because its token percentage value is 51.8%. Topic 2 and topic 3 have a small gap in token percentage value where the token percentage value of topic 2 is 24.5%, and the token percentage value of topic 3 is 23.7%. It can be inferred from these topics that the first topic leads to positive sentiment, while the second and third topics lead to news about vaccination activities and health protocols wherein the process of labeling documents like this will be labeled neutral. This data fits with the fact that the number of tweets with positive and neutral sentiments is very dominant when compared to the number of tweets with negative sentiments.

Fig. 8 shows a visualization of the topic distribution data in general. From the figure, it can be noticed that each topic is well distributed, has quite a distance, and does not overlap each other. On the right side of Fig. 8, 30 most prominent words in the corpus can be identified along with the overall term frequency scale. The term that has the highest frequency is the word "covid," followed by the word “vaksinasi” in the second position. This is natural because this corpus discusses covid and vaccination.

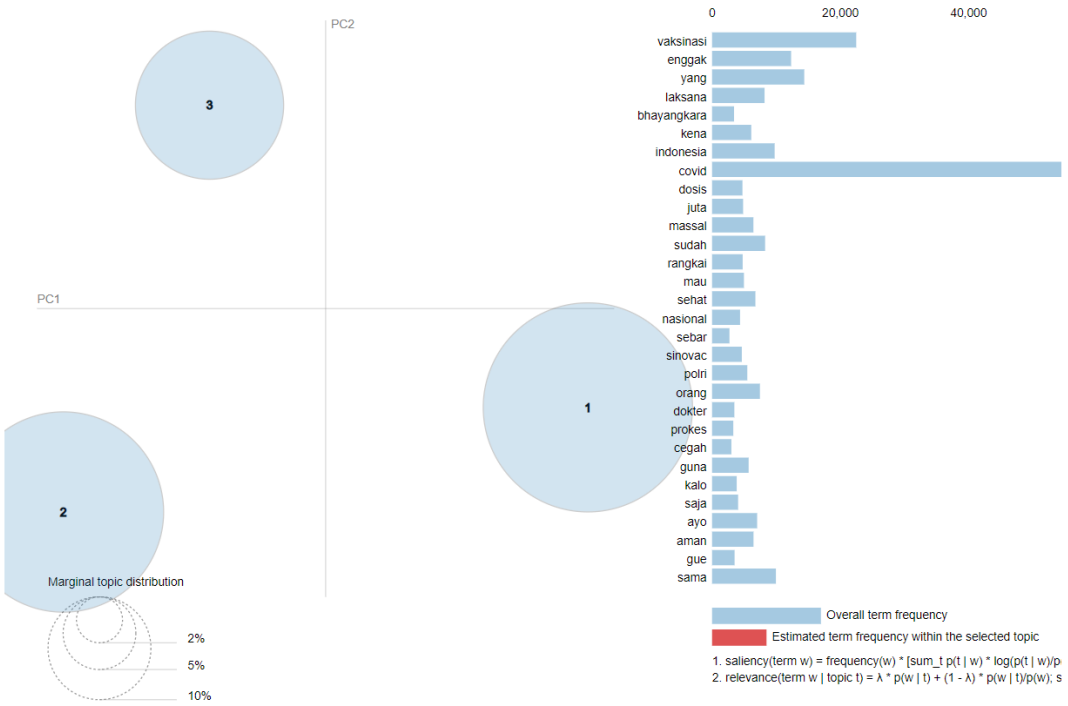

Fig. 8. Topic distribution visualization

3.4. Sentiment Classification

In the initial classification process, Early Stopping was implemented to monitor the training process so that overtraining did not occur. The results of the training can be seen in Table 3 and Fig. 9. From Table 3, it can be identified that training of all scenarios was stopped in the 7th iteration because, in this iteration, the validation loss training started to increase. In general, the training accuracy of the model that used GloVe word embedding is slightly higher than the model that used Fasttext word embedding. On the contrary, the test accuracy of the model that used Fasttext word embedding is slightly higher than the model that used GloVe word embedding. This is in line with research [39], where the model that used fasttext produces higher accuracy than the model that used word embedding Glove on the Indonesian news dataset. In addition, the use of a slang word dictionary for preprocessing data can improve training and validation of model accuracy for both models that used Fasttext and GloVe word embedding. However, the usage of slang word preprocessing did not increase the test accuracy of the model, which used GloVe word embedding.

Table 3. Training results with Early Stopping Monitoring

Scenario	Embedding	Slang Word Preprocessing	Number of Epoch	Training Accuracy	Validation Accuracy	Test Accuracy
1	Fasttext	No	7	78.97%	73.92%	74.70%
2	Fasttext	Yes	7	79.13%	74.35%	75.13%
3	Glove	No	7	79.58%	74.00%	74.44%
4	Glove	Yes	7	79.68%	74.01%	73.50%
Existing Method (without Early Stopping Monitoring)						
Naïve-Bayes	TF-IDF	No	-	-	-	43.99%
Naïve-Bayes	TF-IDF	Yes	-	-	-	42.68%

Training processes in all scenarios were stopped by Early Stopping at a quite low number of training iterations. It is possible that the performance of the model can still increase in more iterations of training, even though it would fluctuate. Therefore, the monitoring of the training process was replaced with the Modelcheckpoint. Authors [40] stated that Modelcheckpoint could improve the Bidirectional LSTM model in a large number of iteration training. In this study, the Modelcheckpoint would monitor the validation accuracy during training and store the model which achieved the highest validation accuracy. In order to minimize the possibility of overfitting, the Dropout layer parameter value implemented in the model was changed from the previously configured 0.3 to 0.5. In addition, the L2 regularizer with parameter 0.01 was also implemented in

the model with the same purpose. All scenarios were trained up to 100 iterations. The results of the training can be found in Table 4 and Fig. 10.

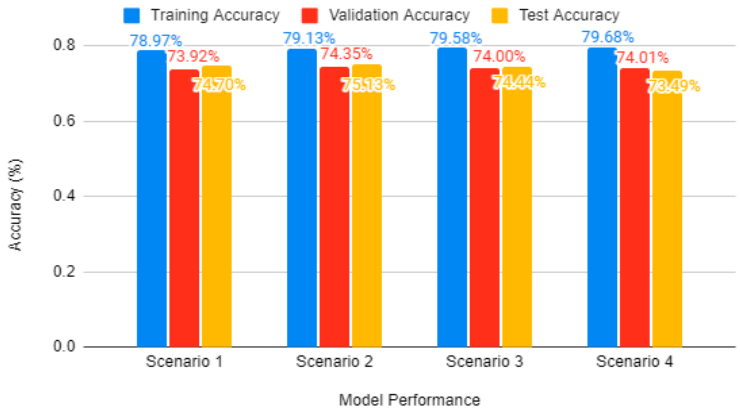

Fig. 9. Model Performance with Early Stopping monitoring

Table 4. Training result with Modelcheckpoint Monitoring

Scenario	Embedding	Slang Word Preprocessing	Number of Epoch	Training Accuracy	Validation Accuracy	Test Accuracy
5	Fasttext	No	96	77.17%	75.23%	75.77%
6	Fasttext	Yes	53	76.94%	75.11%	75.43%
7	Glove	No	92	75.98%	74.11%	74.70%
8	Glove	Yes	63	75.81%	74.49%	74.30%
Existing Method (without Modelcheckpoint Monitoring)						
Decision Tree	TF-IDF	No	-	-	-	70.88%
Decision Tree	TF-IDF	Yes	-	-	-	70.45%

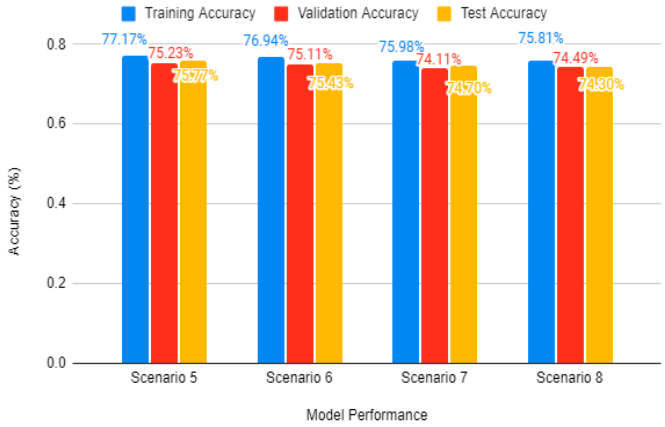

Fig. 10. Models Performance with Modelcheckpoint monitoring

From Table 4, it can be identified that the performance of the model can still improve in several training iterations. The significant difference between the training results monitored using Early Stopping in Table 3 and those monitored using the Modelcheckpoint in Table 4 is the gap between training accuracy and validation accuracy or test accuracy. In scenarios 1 to 4, the accuracy gap was about 5%, while in scenarios 5 to 8, the gap was only 1% to 2%. On the other hand, the validation accuracy and test accuracy performance in scenarios 5 to 8 were higher than in scenarios 1 to 4. In scenarios 5 to 8, the validation accuracy and test accuracy models using Fasttext embedding were slightly higher than those using GloVe embedding. On the other hand, the usage of a slang word dictionary as preprocessing did not increase the accuracy of the model in scenarios 5 to 8.

4. CONCLUSION

In this study, data were collected on discussions of Indonesian people about vaccines on Twitter. The scraping process resulted in 262306 tweets. The data were filtered and cleaned, leaving 83384 tweets. It can

be concluded from the data that although there are people who have negative sentiments towards vaccines which is only 8%, the acceptance of the Indonesian people towards vaccines is quite high is 42%. Meanwhile, 50% sentiment of the discussion is neutral. Public discussion about vaccines began in September 2020. The highest number of tweets appeared in January 2021, which was 23492 tweets.

Based on the measurement results using the Coherence Score, dividing the data into 3 topics resulted in the highest Coherence Score, which was 0.4824. The first topic has a token percentage value of 51.8%, leading to positive sentiment, while the second and third topics lead to news about vaccination activities and health protocols which are neutral, with a token percentage value of 24.5% and 23.7%, respectively.

In general, the performance of the Bidirectional LSTM model in this research only reached around 73% - 75%, even with various scenarios. The highest test accuracy test was generated by a model that used Fasttext word embedding.

The usage of slang words could not increase the test accuracy in this study. The use of the Modelcheckpoint to monitor model performance during training could produce a model with a slightly higher test accuracy compared to a model whose performance was monitored using Early Stopping.

In the next study, we plan to deal with the imbalance problem in the data collected. In addition, topic modeling also needs to be conducted on data of certain months to understand in more detail what topics are discussed at certain times. Lastly, this research has not tried to use a transformer-based model, which is the current state-of-the-art word embedding.

REFERENCES

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, 2020, https://doi.org/10.1016/S0140-6736(20)30183-5.

[2] H. Xu, C. Yan, Q. Fu, K. Xiao, Y. Yu, D. Han, W. Wang, J. Cheng, "Possible environmental effects on the spread of COVID-19 in China," *Sci. Total Environ.*, vol. 731, 2020, https://doi.org/10.1016/j.scitotenv.2020.139211.

[3] S. Olivia, J. Gibson, and R. Nasrudin, "Indonesia in the Time of Covid-19," *Bull. Indones. Econ. Stud.*, vol. 56, no. 2, 2020, https://doi.org/10.1080/00074918.2020.1798581.

[4] M. E. Halloran, I. M. Longini, and C. J. Struchiner, "Design and Analysis of Vaccine Studies: Introduction," *Des. Anal. Vaccine Stud.*, vol. 36, 2009, https://doi.org/10.1007/978-0-387-68636-3.

[5] T. Thanh Le, Z. Andreadakis, A. Kumar, R. Gómez Román, S. Tollefsen, M. Saville, S. Mayhew, "The COVID-19 vaccine development landscape," *Nature reviews. Drug discovery*, vol. 19, no. 5. 2020, https://doi.org/10.1038/d41573-020-00073-5.

[6] I. D. A. P. Dwipayana, "Efforts in Securing Vaccine for Covid-19 Outbreak in Indonesia," *Heal. Notions*, vol. 4, no. 10, 2020, https://doi.org/10.33846/hn41003.

[7] D. Susilo, T. D. Putranto, and C. J. S. Navarro, "9 Performance of Indonesian Ministry of Health in Overcoming Hoax About Vaccination Amid the COVID-19 Pandemic on Social Media," *Nyimak J. Commun.*, vol. 5, no. 1, 2021, https://doi.org/10.31000/nyimak.v5i1.4100.

[8] "Twitter.Inc," 2021. Twitter.com.

[9] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, "Twitter and Research: A Systematic Literature Review through Text Mining," *IEEE Access*, vol. 8, 2020, https://doi.org/10.1109/ACCESS.2020.2983656.

[10] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer," pp. 122–127, 2021, https://doi.org/10.1109/EIConCIT50028.2021.9431852.

[11] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid - 19 Menggunakan Algoritma Naïve Bayes Classifier," *Jurnal Data Mining dan Sistem Informasi*, vol. 2, no. 2, pp. 1–9, 2021, https://ejurnal.teknokrat.ac.id/index.php/JDMSI/article/view/1344.

[12] R. A. Habsi, R. A. D. Anggoro, M. A. Valio, Y. Widiastiwi, N. Chamidah, "Analisis Sentimen Terhadap Vaksin Covid-19 di Jejaring Sosial Twitter Menggunakan Algoritma Naïve Bayes," *Seminar Ilmiah Nasional Online Mahasiswa Ilmu Komputer dan Aplikasinya*, vol. 2, no. 2, pp. 239–248, 2021, https://conference.upnvj.ac.id/index.php/senamika/article/view/1714.

[13] M. Lestandy, A. Abdurrahim, and L. Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naive Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, 2021, https://doi.org/10.29207/resti.v5i4.3308.

[14] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than Bags of Words: Sentiment Analysis with Word Embeddings," *Commun. Methods Meas.*, vol. 12, no. 2–3, 2018, https://doi.org/10.1080/19312458.2018.1455817.

[15] A. R. W. Rapsanjani and E. Junianto, "Implementasi Probabilistic Neural Network Dan Word Embedding Untuk Analisis Sentimen Vaksin Sinovac," *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 2, 2021, https://doi.org/10.51977/jti.v3i2.588.

[16] G. A. Sandag, A. M. Manueke, and M. Walean, "Sentiment Analysis of COVID-19 Vaccine Tweets in Indonesia Using Recurrent Neural Network (RNN) Approach," in *2021 3rd International Conference on Cybernetics and*

Intelligent System (ICORIS), 2021, pp. 1–7, https://doi.org/10.1109/ICORIS52787.2021.9649648.
[17] JustAnotherArchivist, "snscrape," 2021. https://github.com/JustAnotherArchivist/snscrape (accessed Sep. 29, 2021).
[18] K. K. Agustiningsih, E. Utami, and H. Al Fatta, "Sentiment Analysis of COVID-19 Vaccine on Twitter Social Media: Systematic Literature Review," in *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2021, pp. 121–126, https://doi.org/10.1109/ICITISEE53823.2021.9655960.
[19] M. Nazief, B. A. A. & Adriani, "Confix- stripping: Approach to Stemming Algorithm for Bahasa Indonesia," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 38, no. 4, 2005.
[20] E. Utami, I. Oyong, S. Raharjo, A. Dwi Hartanto, and S. Adi, "Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia," *Appl. Comput. Informatics*, vol. ahead-of-p, no. ahead-of-print, Jan. 2021, https://doi.org/10.1108/ACI-03-2021-0054.
[21] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 226–229, https://doi.org/10.1109/IALP.2018.8629151.
[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.
[23] S. Poria, A. Hussain, and E. Cambria, *Multimodal Sentiment Analysis*. Springer Nature, 2018, https://doi.org/10.1007/978-3-319-95020-4.
[24] T. T. Mengistie and D. Kumar, "Deep Learning Based Sentiment Analysis On COVID-19 Public Reviews," in *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2021, pp. 444–449, https://doi.org/10.1109/ICAIIC51459.2021.9415191.
[25] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017, https://doi.org/10.1162/tacl_a_00051.
[26] A. G. D'Sa, I. Illina, and D. Fohr, "BERT and fastText Embeddings for Automatic Detection of Toxic Speech," *Proc. 2020 Int. Multi-Conference Organ. Knowl. Adv. Technol. OCTA 2020*, 2020, https://doi.org/10.1109/OCTA49274.2020.9151853.
[27] Joydeep Bhattacharjee, *FastText Quick Start Guide*. Packt Publishing, 2018.
[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com.
[29] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," *J. Infect. Public Health*, vol. 14, no. 10, 2021, https://doi.org/10.1016/j.jiph.2021.08.010.
[30] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015, https://doi.org/10.48550/arXiv.1412.6980.
[31] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning - whole book," *Nature*, vol. 521, no. 7553, 2016, https://doi.org/10.1038/nature14539.
[32] C. I. Redaksi, "Hot Issue: Jokowi Tunjuk Menkes Baru, Terawan Diganti BGS?," 2020. https://www.cnbcindonesia.com/news/20201221121604-4-210539/hot-issue-jokowi-tunjuk-menkes-baru-terawan-diganti-bgs (accessed Jan. 02, 2022).
[33] "Worldometer," 2021. https://www.worldometers.info/coronavirus/country/indonesia/ (accessed Dec. 29, 2021).
[34] C. I. Redaksi, "Joe Biden Disuntik Vaksin Covid-19," 2020. https://www.cnbcindonesia.com/news/20201222145043-8-210917/joe-biden-disuntik-vaksin-covid-19 (accessed Jan. 02, 2022).
[35] BPMI Setpres, "Presiden Jokowi Menerima Vaksin Covid – 19 Perdana," 2021. https://www.presidenri.go.id/siaran-pers/presiden-jokowi-menerima-vaksin-covid-19-perdana/ (accessed Jan. 02, 2022).
[36] Redaksi Sehat Negeriku, "Pelaksanaan Vaksinasi COVID-19 di Indonesia Membutuhkan Waktu 15 Bulan - Sehat Negeriku," *Kementerian Kesehatan RI*, 2021. https://sehatnegeriku.kemkes.go.id/baca/umum/20210103/2536122/pelaksanaan-vaksinasi-covid-19-indonesia-membutuhkan-waktu-15-bulan/ (accessed Jan. 02, 2022).
[37] A. Sofa, "Covid-19 Varian Delta dan Hal-hal yang Harus Kamu Perhatikan," *Jakarta Smart City*, 2021. https://smartcity.jakarta.go.id/blog/758/covid-19-varian-delta-dan-hal-hal-yang-harus-kamu-perhatikan (accessed Jan. 02, 2022).
[38] R. Oktari, "Bahan Baku Vaksin COVID-19, November Diterima," 2021. https://indonesiabaik.id/infografis/bahan-baku-vaksin-covid-19-november-diterima (accessed Jan. 02, 2022).
[39] R. Adipradana, B. P. Nayoga, R. Suryadi, and D. Suhartono, "Hoax analyzer for indonesian news using rnns with fasttext and glove embeddings," *Bull. Electr. Eng. Informatics*, vol. 10, no. 4, 2021, https://doi.org/10.11591/eei.v10i4.2956.
[40] O. M. A. Alsyaibani, E. Utami, and A. D. Hartanto, "An Intrusion Detection System Model Based on Bidirectional LSTM," in *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2021, pp. 1–6, https://doi.org/10.1109/ICORIS52787.2021.9649612.

BIOGRAPHY OF AUTHORS

Kartikasari Kusuma Agustiningsih is a master's degree student at Universitas Amikom Yogyakarta. She is a graduate of the Universitas Negeri Malang, majoring in Informatics Engineering Education. Now she is working as a teacher in a Vocational High School. Email: kartikasarikusuma@students.amikom.ac.id

Ema Utami is a senior lecturer at Universitas Amikom Yogyakarta. She is a professor of computer science who graduated from Gajah Mada University. Ema specializes in Artificial Intelligence, Databases, NLP, and Data Science. Email: ema.u@amikom.ac.id

Omar Muhammad Altoumi Alsyaibani is a master's degree student at Universitas Amikom Yogyakarta. He is a graduate of the Universitas Negeri Yogyakarta, majoring in Informatics Engineering Education. Now, he is working as a teacher in a Vocational High School. Email: omar@smkn2banjarbaru.sch.id