Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 5, No. 2, pp. 126-134, 2024

Implementation of Bidirectional Long Short Term Memory (BiLSTM) Algorithm with Embedded Emoji Sentiment Analysis of Covid-19 Anxiety Level and Socio-Economic Community

Jenie Marcelina[1*], Eneng Tita Tosida[2], Adriana Sari Aryani[3]

[1,2,3]Department of Computer Science, Universitas Pakuan
Jl. Pakuan PO Box 452, 16143 Bogor, Jawa Barat, Indonesia
**Corresponding author email: jenniemarcelina@gmail.com*

Abstract

The COVID-19 pandemic has presented multidimensional challenges in Indonesia, significantly affecting social, economic, and public health at the level of anxiety. Public anxiety related to the pandemic can be reflected in online media, especially Twitter, which is the main channel for information sharing and emotional expression. This study aims to understand the level of public anxiety in relation to the aftermath of the COVID-19 pandemic by using a classification method. Classification is carried out using the Knowledge Discovery in Database (KDD) method with the Bidirectional LSTM algorithm and emoji embedding sentiment analysis, and K-Fold Cross Validation testing is also carried out with various optimizers. The final result of the best accuracy rate obtained was 98.08%. This shows that the classification model created is good.

Keywords: Public Anxiety, Covid-19, Emoji embedded, Bidirectional LSTM, Sentiment analysis

1. Introduction

This study aims to understand the impact of the COVID-19 pandemic on the social, economic and public health situation in Indonesia. The pandemic has created unprecedented challenges, such as mobility restrictions that have impacted the health, social and economic crisis. Surveys conducted by Unicef and BPS show that the socio-economic conditions of communities are severe, with many households losing their jobs. This also affected children and adolescents, who experienced learning difficulties and limited social interaction. The pandemic created a psychological burden with increased levels of anxiety, fear and mental distress (Chen et al., 2018).

Sentiment analysis of public opinion on online media, especially Twitter, is the focus of this research. A document classification method based on key words that reflect positive, negative and neutral sentiments will be used to understand the dynamics of public anxiety during the COVID-19 pandemic (Hung et al., 2020; Tosida et al., 2021). Previous research shows that methods such as LSTM and BiLSTM are effective in sentiment analysis, while emoji embedding can also provide additional information in understanding the emotion of the text. In general, the function of emoji is to convey emotional expressions that cannot be expressed through text alone. Compared to words, emojis can describe one's feelings more accurately and directly (Jiawei et al., 2022). When examining informal texts such as tweets, blogs or comments, emojis provide important clues about the user's feelings. The widespread presence of emoji provides new opportunities for us to study the expression of feelings in the context of writing. This research seeks to use emojis that reflect emotions as a factor influencing sentiment on social media platforms.

In further research, it will focus on the Implementation of the BiLSTM Algorithm with Embedded Emoji Sentiment Analysis on the Level of COVID-19 Anxiety and Socio-Economic Society. The ultimate goal is to gain a deeper understanding of people's anxiety levels and the socio-economic impacts faced during the pandemic, using the latest sentiment analysis and data processing techniques.

2. Literature Review

2.1 Anxiety Levels

According to Arifiyanti & Wahyuni, (2020), anxiety is a natural response to threatening situations and can occur to anyone in the face of change, new experiences, and in the search for identity and meaning in life. Excessive anxiety can disrupt the balance of one's life. WHO, as a world health agency, emphasizes the importance of overall health, including mental and social health, in addition to physical. This suggests the need to address anxiety arising from the COVID-19 pandemic to ensure the overall well-being of society (Liu et al., 2021). Anxiety is one of the most common mental health disorders, often alongside depression. Research in May 2020 found that most individuals, especially in the 16-24 age range, experience symptoms of anxiety and depression caused by various factors such as financial, academic, and loneliness problems (Kaligis et al., 2021; Rumelhart et al., 1985; Saxena & Sukumar, 2018).

2.2 Emoji Embedding

Social media such as Twitter is the focus of exploration due to its massive usage, generating content with user sentiment and widely used emojis. Emoji are character images commonly used in electronic messages and social media posts, which can express emotions. Sentiment analysis on social media is becoming important for governments and organizations, increasing attention on emoji prediction for emotion analysis. Emoji embedding methods can improve analysis accuracy by extracting the semantics and emotions of words (Tavan et al., 2020). Although there are thousands of emoji types, only a small number of them are used in this study, which can be seen in Table 1.

Table 1: Emoji Examples and Their Label Name Descriptions

No.	Emoji	Deskripsi
1.	😭	loudly crying face
2.	😊	smiling face
3.	😂	face with tears of joy
4.	😠	angry face
5.	☹ □	frowning face
…	…	…
100.	🙏	folded hands
…	…	…
223.	❤ □	red heart

2.3 Bidirectional Long Short-Term Memory

Long Short-Term Memory (LSTM) is a development of Recurrent Neural Network (RNN) architecture designed to overcome the Vanishing Gradient problem that often occurs in RNN. LSTM uses block memory cells with gate mechanisms, including forget gate, input gate, and output gate, to capture long term dependencies in long sequential data. The single neuron structure of LSTM can be seen in Figure 1.

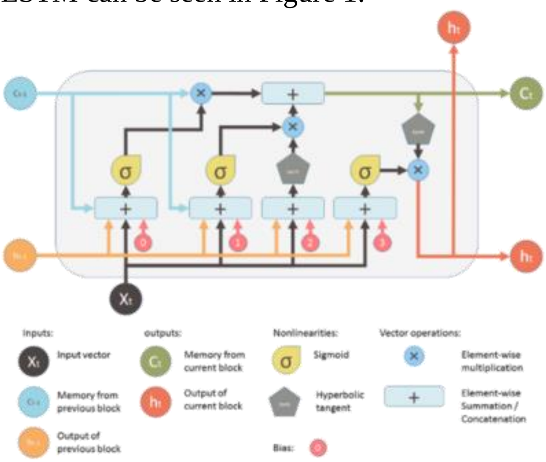

Figure 1: Illustration of LSTM

LSTM has two main parts, the cell state and the hidden state, where the cell state is the main network for data flow that remains stable with the possibility of linear transformation. Data can be modified through sigmoid gates that resemble matrix operations with individual weights. LSTM is designed to overcome the long-term dependency problem by using specific calculation formulas for each layer.

a) *Forget Gate*

$$f_t = \sigma\left(W_f \, . \, [h_{t-1}, X_t] + b_f\right) \tag{1}$$

b) *Input* Gate

$$i_t = \sigma\left(W_i \, . \, [h_{t-1}, X_t] + b_i\right) \tag{2}$$

c) *Memory Update*

$$\tilde{C}_t = tanh(W_c \, . \, [h_{t-1}, X_t] + b_c) \tag{3}$$

$$C_t = f_t \, . \, c_{t-1} + i_t \, . \, c_t \tag{4}$$

d) *Output* Gate

$$O_t = \sigma\left(W_o \, . \, [h_{t-1}, X_t] + b_o\right) \tag{5}$$

$$h_t = O_t \, . \, tanh(C_t) \tag{6}$$

The formulas in the LSTM layer use sigmoid activation functions that describe internal activity relationships in a linear or non-linear manner, to determine the activation of neurons. Common activation functions include sigmoid, tanh, and softmax. BiLSTM is an extension of LSTM that overcomes the drawback of reading data from one direction only. With two layers that move in opposite directions, i.e. forward and backward, BiLSTM can understand the context of the previous and following words in more depth, improving the understanding of patterns in sentences. This model uses two separate LSTMs for the forward and backward directions, providing a more holistic perspective on the data. The structure of the BiLSTM algorithm can be seen in Figure 2.

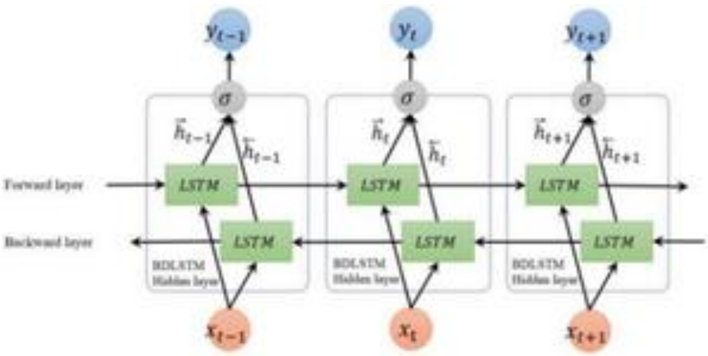

Figure 2: Illustration of BiLSTM

Each hidden unit output unit in the lower and upper layers is combined to form a feature value of the word with a longer size than using a regular LSTM. Because the feature value is longer, the information that will be processed in the next stage, namely feed forward, will classify more accurately.

2.4 *K-Fold Cross Validation*

Cross Validation is a technique in data mining that divides data into training and test data to achieve optimal accuracy. One commonly used method is K-Fold Cross Validation, where the dataset is divided into K partitions randomly and experimented K times, using one partition as test data and the other as training data.

3. Materials and Methods

3.1. Materials

Twitter data was taken through scrapping tweets regarding the level of Covid-19 Anxiety and Socio-Economic Society with keywords related to social, economic, and health aspects. The data was taken from social focus (social, PSBB, interaction, stigma, online), economy (economy, PPKM, recession, inflation), and health (mental health, anxiety, covid-19, vaccine, health services) from December 1-31, 2022. Total data of 3,707 tweets, with 223 emoji types out of 3,570 Unicode emoji types.

3.2. Methods

KDD is a complex process of discovering new useful patterns in data. This study uses Twitter data, through the Preprocessing stage, word embedding with one hot encode, BiLSTM classification, and testing using confusion matrix.

a) Data Collection Stage

Data from Twitter was crawled using keywords in three aspects: social, economic, and health. Aspects were chosen based on topics that are often discussed on Twitter. Interviews and validation with Indonesian language experts were conducted. The crawling results were saved in a csv file. Then all the data was put together in related aspects and validated by experts, then saved in .csv format with the amount of data for each keyword shown in Table 3.

b) Data Cleaning and Data Integration

Data cleaning and integration were performed to remove noise and combine data from Twitter and news portals. Data was collected by scrapping using the tool on Anaconda Navigator for social, economic and health aspects. Results can be seen in Table 2, which shows the amount of data based on each aspect in the study.

Table 2: Description of the Amount of Data Based on Each Aspect in Research

Aspect	Keywords	Sum
Social	social, PSBB, interaction, stigma, and online.	1,115
Economic	economy, PPKM, recessions, inflation.	1,288
Health	mental health, anxiety, covid-19, vaccines and healthcare.	1,304
	Total	3,707

Initial data often has irrelevant attributes and noise. Before the classification process, the data needs to go through preprocessing to adjust the format, remove irrelevant characters, and simplify the vocabulary. Preprocessing stages include converting emojis, removing usernames, hashtags, URLs, and mentions, removing punctuation, changing lowercase letters, removing stopwords, stemming, and translating data into English. These stages can be seen in Figure 3.

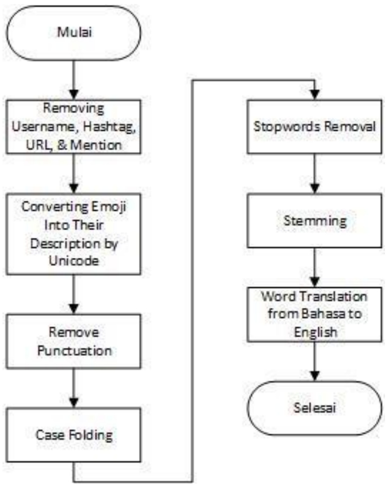

Figure 3: Preprocessing Stage

The text noise removal process includes usernames, hashtags, URLs, mentions, punctuation marks, and other non-alphanumeric characters. In this study, emojis are also considered, converted into labels based on the Unicode dictionary before removing noise and non-alphanumeric characters. There are 3,570 types of emojis in the Unicode dictionary that are converted into label names or their specific descriptions, as well as into labels. Emojis are manually identified and extracted by converting them into label names or descriptions based on the Unicode dictionary. This is done so that the emoji are well integrated in the text data and are not deleted during the subsequent non-alphanumeric character cleaning process. This is important because some emojis cannot be grouped into specific sentiment categories. Examples of emoji in the data that have been converted into label names or descriptions can be seen in Table 3.

Table 3: Emoji Types Found in Health Aspects

No.	Emoji	Emoji Name/Description Label	Number of Emojis
1	😭	loudly crying face	138
2	😂	face with tears of joy	97
3	🤣	rolling on the floor laughing	89
…	…	…	…

10	😃	grinning face	48
…	…	…	…
100	😞	dissapointed face	1
…	…	…	…
138	🙇	person bowing	1

c) Data Selection and Data Transformation

Data from the database is selected only those that are suitable for analysis. Data transformation involves converting the data to a format that can be processed in data mining, such as with word embedding to convert letters in sentences into vectors. This is important because LSTM cannot process letters, and the output is a value used for classification. This stage is the word weighting stage using One Hot Encoding to convert words into numeric vectors. This example selects the 2nd data as a sample that is converted into a vector with One Hot Encoding, where one element is worth 1 according to the word in the dictionary. This is important because BiLSTM requires input in the form of vectors, not letters, to process data with the method.

d) Data Mining Process

Data mining is the process of finding interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The selection of the right method or algorithm depends on the objectives and the overall KDD process. In the data mining process, the Bidirectional LSTM (BiLSTM) algorithm is used. Technically, BiLSTM processing applies two separate LSTMs, one for the forward layer and one for the backward layer. The application of two separate LSTMs will then be combined using the BiLSTM output equation and then the latest BiLSTM output value is reprocessed using the softmax activation function. The BiLSTM output equation follows:

$$y_t = W_{\overrightarrow{hy}} h_{\vec{t}} + W_{hy^{\leftarrow}} h_{t^{\leftarrow}} + b_y$$

e) Pattern Evaluation

After applying BiLSTM, an evaluation is performed to identify interesting patterns and check the success of the prediction by calculating the accuracy using confusion matrix. This research focuses on the accuracy parameter as an assessment of the success of the method.

4. Results and Discussion

Analysis of data from Twitter social media on anxiety about the Covid-19 pandemic, which is sorted into three main topics, namely health, social, and economic, resulted in the following classification:

4.1 Classification Results

4.1.1 Health Aspect Sentiment Classification Results

Health sentiment analysis with emoji on 1,304 data shows: 571 negative, 516 positive, and 217 neutral can be seen in Figure 4 the level of negative anxiety regarding health in the Covid-19 pandemic is the most discussed by the public, as seen from the data in Figure 4. The average Indonesian from December 2022 showed more anxiety in the health aspect, with 571 negative, 199 positive, and 74 neutral sentiments out of 431 test data. This represents the dominance of negative anxiety (46%) over positive (37%) and neutral (17%) anxiety over the two years since December 2022.

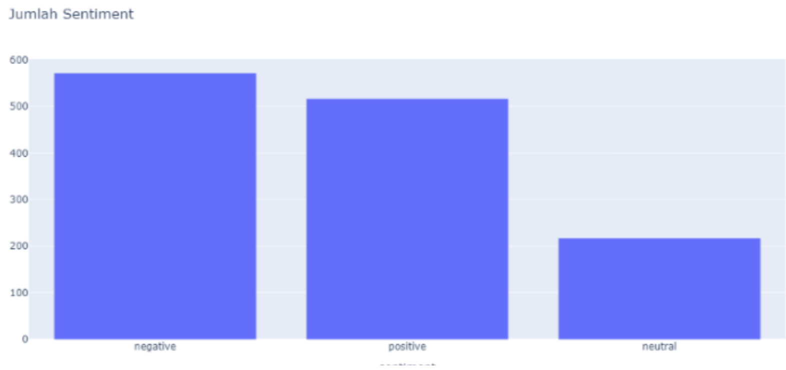

Figure 4: Results of Health Aspect Sentiment Classification with Emoji

4.1.2 Social Aspect Sentiment Classification Results

The results of sentiment analysis of social aspects with 1,115 data show the dominance of negative (490 data), neutral (371 data), and positive (254 data) anxiety can be seen in Figure 5. In the Covid-19 pandemic, Indonesians tend to discuss more negative social anxiety, as seen from the average in December 2022: 128 negative, 118 neutral, and 81 positive out of 391 test data. Negative (33%) and neutral (31%) social anxiety were higher than positive anxiety (21%) for two years after December 2022.

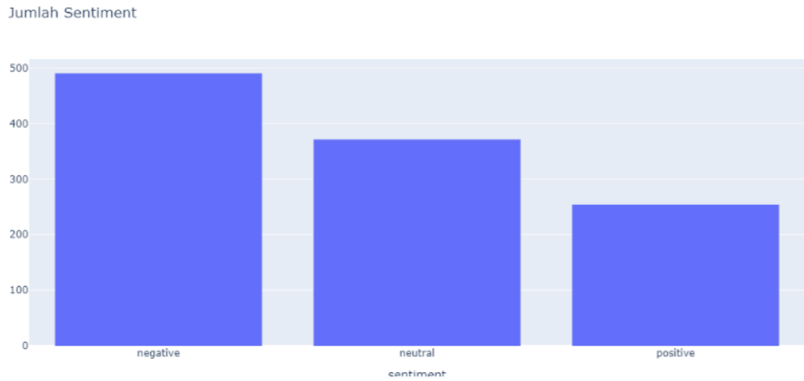

Figure 5: Results of Social Aspect Sentiment Classification with Emoji

4.1.3 Economy Aspect Sentiment Classification Results

The results of the economic sentiment analysis of 1,288 data show the tendency of the dominance of negative (487 data), positive (401 data), and neutral (400 data) anxiety can be seen in Figure 6. In the Covid-19 pandemic, Indonesians discuss more negative economic anxiety, as seen in Figure 5 from the average in December 2022: 136 negative, 114 positive, and 118 neutral from 368 test data. Negative economic anxiety (37%) is higher than positive (31%) and neutral (32%) for a period of two years after December 2022.

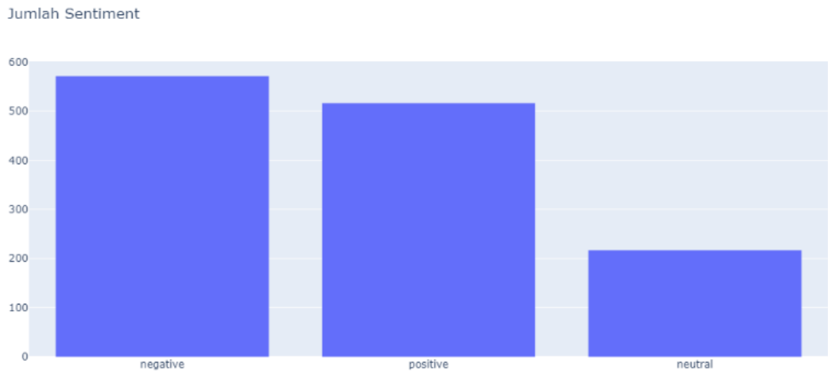

Figure 6: Results of Economy Aspect Sentiment Classification with Emoji

4.2 Interpretation of Knowledge

The interpretation of the knowledge generated to increase public awareness of mental health, especially anxiety in the post-19 pandemic time to be more vigilant about the health of the environment is as follows:

1. Based on Figures 4, 5, and 6, it shows that on average, people tend to dominate feeling negative anxiety consisting of health aspects, social aspects, and economic aspects since the 2-year period of the covid-9 pandemic passed.

2. From the classification results in Figures 4, 5, 6, there are still many people who ignore their mental health. Therefore, it is necessary to increase awareness of the importance of maintaining mental health, especially anxiety.

4.3 Results Evaluation

Confusion matrix testing was conducted to see the results of BiLSTM prediction and classification tests on each aspect with or without emoji embedding. The overall results of classification validation are shown in the confusion matrix in Figures 7, 8, and 9 for each aspect.

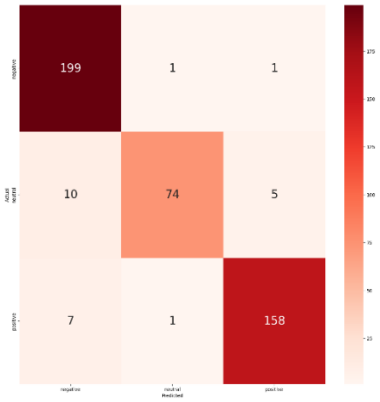

Figure 7: Confusion Matrix Results of Health Aspects with Emoji

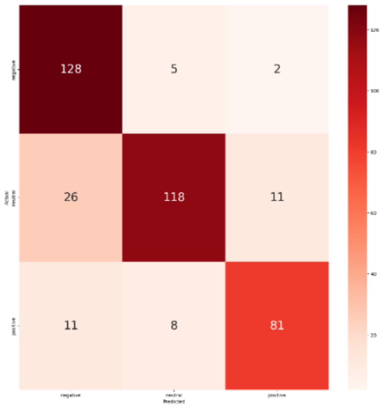

Figure 8: Confusion Matrix Result of Social Aspect with Emoji

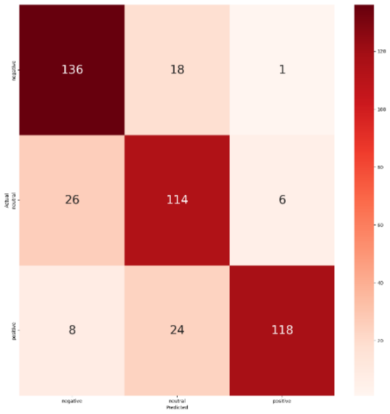

Figure 9: Confusion Matrix Result of Social Aspect with Emoji

The confusion matrix results in Figures 7, 8, and 9 show the dominance of negative sentiments in all aspects, with the highest accuracy. Tests with emoji embedding achieved 94.52% accuracy for Health Anxiety, 83.85% for Social Anxiety, and 81.60% for Economic Anxiety. Without emoji, the accuracy dropped to 76% for Health Anxiety, 58% for Social Anxiety, and 65% for Economic Anxiety. The average accuracy of both tests is in Table 4.

Table 4: Results of Average Accuracy Testing

No	Aspect with Emoji	Accuracy	Aspect without Emoji	Accuracy
1	Health Aspect	94.52%	Health Aspect	76%
2	Social Aspect	84.85%	Social Aspect	58%
3	Economic Aspect	81.60%	Economic Aspect	65%
	Average	**86,99%**	**Average**	**66.33%**

The number of positive, negative, and neutral reviews in each aspect of the confusion matrix is obtained, with an average accuracy value of 86.99% with emoji and 66.33% without emoji. It can be concluded that BiLSTM with emoji classifies sentiment well, has high accuracy and few errors. Testing all aspects showed good results with Adam's optimizer without parameter tuning. However, the data for each aspect with or without emoji was tested again using k-fold cross-validation and manual learning rate settings with Adam's optimizer (0.01) and SGD (0.01 learning rate and 0.9 momentum). The accuracy of the results of k-fold cross-validation testing using BiLSTM can be seen in Table 5.

Table 5: Data Accuracy Results with Emoji Aspects K-Fold Cross Validation Testing

No.	Aspect with Emoji	Accuracy with Adam	Accuracy with K-Fold Cross Validation + Adam: Learning Rate 0.01	Accuracy with K-Fold Cross Validation + SGD: Learning Rate 0.01
1.	Health Aspect	94.52%	99.29%	63.12%
2.	Social Aspect	84.85%	97.10%	45.64%
3.	Economic Aspect	81.60%	97.85%	65.59%
	Average	**86.99%**	**98.08%**	**58.11%**

While the dataset without emoji embedding obtained the following results with k-fold cross validation testing, can be seen in Table 6.

Table 6: Data Accuracy Results with Aspects Without Emoji K-Fold Cross Validation Testing

No.	Aspect with Emoji	Accuracy with Adam	Accuracy with K-Fold Cross Validation + Adam: Learning Rate 0.01	Accuracy with K-Fold Cross Validation + SGD: Learning Rate 0.01
1.	Health Aspect	76%	95.39%	45,23%
2.	Social Aspect	58%	96.30%	46%
3.	Economic Aspect	65%	96.40%	46%
	Average	**66,33%**	**96.03%**	**45.74%**

Testing k-fold cross-validation and tuning the parameters of the BiLSTM showed a significant improvement in accuracy. From Table 19, it can be seen that by tuning the Adam optimizer parameters, the accuracy increased dramatically from 86.99% to 98.08% with emoji and from 66.33% to 96.03% without emoji. However, tuning the parameters of the SGD optimizer did not result in a significant improvement from the initial accuracy without k-fold cross-validation and learning rate determination. This is due to various factors including dataset size and problem characteristics. In the whole, the use of the Adam optimizer with a tuning parameter of 0.01 is more effective in dealing with the sentiment classification problem at the anxiety level.

4.3.1 Wordcloud

The performance patterns are evaluated based on the positive, negative, and neutral results of each aspect of the dataset. Wordcloud displays the most frequently occurring words, taken from the base word after the stemming process. The occurrences of these words are counted according to the positive, negative, and neutral corpus, and visually displayed in the form of a wordcloud. The wordcloud for each aspect with emoji can be seen in Figures 10, 11, and 12.

Figure 10: Health Aspect Wordcloud with Emoji

Figure 11: Social Aspect Wordcloud with Emoji

Figure 12: Economic Aspect Wordcloud with Emoji

A wordcloud of health aspects with emojis showed a focus on anxiety in Twitter user reviews, especially regarding health during the pandemic. Words such as 'health', 'vaccine', and 'impact' reflect individual awareness, while 'anxiety', 'anxious', 'lockdown', and 'crying' indicate intensity of feeling. The discussion reflects the concern for Indonesia in dealing with health issues, with the 'face' emoji used to reinforce the message. On the social aspect Wordcloud highlights discussions on interaction, stigma, and expression of emotions through emoji. This is reflected in words such as 'stigma', 'interaction', and emoji variations such as 'face', 'loudly', 'frowning', and 'smiling' and the economic wordcloud with emoji shows a focus on words such as 'recession', 'inflation', 'economic', and 'pandemic', highlighting the impact of the pandemic on the Indonesian economy in 2023. The use of the 'face' emoji increases the clarity of the message conveyed.

5. Conclussion

This study compares classification accuracy using BiLSTM before and after applying k-fold cross validation testing and parameter tuning adjustments, resulting in the best results. Adjusting the parameters in Adam's optimizer by setting the learning rate to 0.01 significantly improved the model's performance, especially in terms of accuracy. The average accuracy for aspects involving emoji reached 98.08%, with health aspects reaching 99.29% accuracy, social aspects reaching 97.10%, and economic aspects reaching 97.85%. This finding indicates that Adam's parameter tuning effectively improved the model's performance, especially in handling sentiments with anxiety levels.

The test results show that sentiment classification using the BiLSTM algorithm with k-fold cross validation and parameter tuning on a training dataset integrated with emoji provides better accuracy than sentiment classification without emoji. Although emojis do not directly determine sentiment, their use helps improve the optimization of sentiment prediction on text.

The conclusion of the research shows that by combining the Bidirectional Long Short-Term Memory (BiLSTM) method with emoji embedding in sentiment analysis was successfully applied related to the Covid-19 anxiety level and socio-economic society. Although the tweet data shows a higher proportion of negative sentiments, signaling a still high level of anxiety related to health, social, and economic issues where the Covid-19 pandemic has passed two years. The impact of the pandemic remains significant after its end, and the high anxiety level shows that Covid-19 is not the only factor affecting people's anxiety.

References:

Arifiyanti, A. A., & Wahyuni, E. D. (2020). Emoji And Emoticon In Tweet Sentiment Classification. *Proceeding - 6th Information Technology International Seminar, ITIS 2020, 145–150.* https://doi.org/10.1109/ITIS50118.2020.9320988

Chen, Y., Yuan, J., You, Q., & Luo, J. (2018). Twitter Sentiment Analysis via Bi-sense Emoji Embedding and. *26th ACM International Conference on Multimedia,* 117–125.

Hung, M., Lauren, E., Hon, E. S., Birmingham, W. C., Xu, J., Su, S., Hon, S. D., Park, J., Dang, P., & Lipsky, M. S. (2020). Social network analysis of COVID-19 Sentiments: Application of artificial intelligence. *Journal of Medical Internet Research*, *22*(8). https://doi.org/10.2196/22590

Jiawei Han, Jian Pei, H. T. (2022). Data Mining Concept and Techniques. *Fourth Edi*. https://books.google.co.id/books?hl=id&lr=&id=NR1oEAAAQBAJ&oi=fnd&pg=PP1&dq=KDD+method+by+han+et+al+2022&ots=_N1GUJykt2&sig=x9VbgYu4j9D90cwOMEa0-N4vtM&redir_esc=y#v=onepage&q=KDD method by han et al 2022&f=false

Kaligis, F., Ismail, R. I., Wiguna, T., Prasetyo, S., Indriatmi, W., Gunardi, H., Pandia, V., & Magdalena, C. C. (2021). Mental Health Problems And Needs Among Transitional-Age Youth In Indonesia. *International Journal of Environmental Research and Public Health, 18*(8). https://doi.org/10.3390/ijerph18084046

Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., & Lu, X. (2021). Improving Sentiment Analysis Accuracy With Emoji Embedding. *Journal of Safety Science and Resilience*, *2*(4), 246–252. https://doi.org/10.1016/j.jnlssr.2021.10.003

Rumelhart, David E., Geoffrey E. Hinton, and R. J. W. (1985). Learning internal representations by error propagation.

Saxena, A., & Sukumar, T. R. (2018). Predicting bitcoin price using lstm And Compare its predictability with arima model. *International Journal of Pure and Applied Mathematics*, *119*(17), 2591–2600. https://doi.org/10.13140/RG.2.2.15847.57766

Tavan, E., Rahmati, A., & Ali Keyvanrad, M. (2020). Persian Emoji Prediction Using Deep Learning and Emoji Embedding. *2020 10h International Conference onComputer and Knowledge Engineering, ICCKE 2020*, 350–355. https://doi.org/10.1109/ICCKE50421.2020.9303639

Tosida, E. T., Erniyati, K., & Talib Bon, A. (2021). Sentiment Analysis Using the Support Vector Machine For Community Compliance Representation in The Covid-19 Pandemic Period.

Acknowledgments

Acknowledgments are conveyed to the Computer Science Department of the Faculty of Mathematics and Natural Sciences, Pakuan University.