JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 1, Maret 2023, Pp. 172-180

DEPRESSION DETECTION ON SOCIAL MEDIA TWITTER USING XLNET METHOD

Fika Apriliani*[1], Warih Maharani[2]
1. Telkom University, Bandung, Indonesia
2. Telkom University, Bandung, Indonesia

Article Info
Keywords: Depression, Social Media, Twitter, XLNET, NLP.

Article history:
Received 13 November 2022
Revised 27 November 2022
Accepted 4 December 2022
Available online 1 March 2023

DOI :
https://doi.org/10.29100/jipi.v8i1.3345

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

ABSTRACT
Depression is a serious mental illness. Depression is usually characterized by feelings of sadness, hopelessness, anxiety, restlessness, and even loss of life. However, not everyone who experiences depression can get professional treatment. If depression is left unchecked, it can worsen the mental health conditions experienced by a person. Social media, one of which is the increasingly popular twitter can be utilized to help deal with the problem of undetected mental illness. Based on tweets made by a person twitter social media can be one of the sources to detect depression using the XLNet method. XLNet is one of the NLP (Natural Language Processing) techniques based on machine learning models on text. Based on several tests that have been carried out during the research such as testing various tuning hyper-parameters with different values on the XLNet model, it achieves a good performance value with an average accuracy value of 93.33%.

I. INTRODUCTION

DEPRESSION is a serious mental illness if left without specialized psychological treatment. The World Health Organization (WHO) suggests that depression is a common disease in the world where there are approximately 280 million people experiencing depression with an estimated 3.8% of the affected population including 5.0% of adults and 5.7% of adults over the age of 60 [1]. According to research conducted by YouGov, about 27% of Indonesians have thought about suicide, namely 33% of adults aged 18-24 years and 20% of older adults aged 55 years. In addition, research shows that 36% of Indonesians have committed self-harm including 45% of younger adults and 7% tend to frequently commit self-harm. Research states that 24% of younger adults tend to have mental health disorders [2].

Depression is a disease that is the leading cause of non-fatal disease burden worldwide, accounting for nearly 12% of people living with disability each year (Üstün, Ayuso-Mateos, Chatterji, Mathers, & Murray, 2004) [3]. According to Schreiber, depression is "self-limiting" which means that depression has a tendency that will increase over time even if treatment will take a long time until a person experiences the next depressive episode [3]. As a study conducted by Schreiber related to the identification of depression in women, there are six phases of depression, namely the condition of the self before entering a depression, entering a depressive condition, trying to tell the depressive condition experienced, finding out about oneself and the social world, finding oneself and around the social world environment, and finally being able to see related to oneself and the social world [3]. Depression can be overcome by direct professional care or treatment related to mental health conditions. Many studies have focused only on the depression experienced by a person because depression is under-diagnosed, which is about half of the cases diagnosed by doctors by undergoing primary treatment, and only 13% - 49% receive adequate treatment [4]. However, not all people with depression can receive treatment. As to the results of research conducted by YouGov, about 46% of people are not sure where to get help, 45% are worried about the costs incurred when carrying out treatment, and 33% feel embarrassed to disclose their mental conditions because they are worried about social stigma, and 25% are worried about the time commitment during treatment [2].

Along with the times, the internet is widely used by all people in the world. Internet users increased by 4% in a year as recorded in the Reportal Data Report, which was 4.95 billion internet users in January 2022 from 4.76 billion users in January 2021 [5]. The internet can be utilized in various ways, one of which is social media. Social media users reached 4.2 billion worldwide in January 2021 [6]. Nowadays, social media has become the most important part of a person's life. Social media can be used as a place to share what activities are being carried out and describe the feelings experienced indirectly. Not a few people use social media as a communication medium that can connect with each other. What activities are shared on social media one can easily find out habits and behaviors that can be suspected of forming patterns of behavior of anxiety, depression, stress, or other mental illnesses. One of the most popular social media platforms is Twitter. Currently, Indonesia ranks 6th with the number of twitter social media users reaching 15.7 million users [7]. One of the features of twitter social media is creating tweets. Based on the tweets made by someone on their twitter account, can be used as a source to detect depression.

There have been many studies related to depression detection, such as those conducted by Anu Priya, et al related to the prediction of anxiety, depression, and stress in modern life that utilizes several machine learning algorithms. The results obtained related to depression detection, namely, the Decision Tree algorithm obtained an accuracy of 0, 778 and f1-score of 0, 723. Using the Random Forest algorithm gets an accuracy value of 0, 798 and f1-score of 0, 766. The Naive Bayes algorithm has an accuracy of 0, 855 and an f1-score of 0, 836. The Support Vector Machine (SVM) algorithm obtained an accuracy of 0, 803 and an f1-score of 0, 765. While the K-Nearest Neighbor (KNN) algorithm gets an accuracy value of 0, 721 and an f1-score of 0, 687. Based on this research, f1-score is an important measure in determining the best model. The Naive Bayes algorithm has the highest f1-score value of 0, 836 so it is the best model for the research conducted [8].

Apart from these studies, there is also research conducted by Ahmed Hussaini Orabi, et al on the detection of depression in twitter social media users by utilizing deep learning. The methods used in the research are CNN With Max, Multi-Channel Pooling CNN, Multi-Channel CNN, and BiLSTM (Context-aware attention). Based on the research that has been done, the performance of each model is obtained using 5-fold cross-validation. The optimal results obtained from several methods that have been used can be concluded that the CNN With Max method has the highest accuracy value of 87, 957 and the highest f1-score value of 86, 967 [9].

Another depression detection research was conducted by Yipeng Zhang, et al who monitored depression trends on twitter social media during the COVID-19 pandemic. The algorithms used in the research include Attention BiLSTM, CNN, BERT, RoBERTa, and XLNet using the Chunk-Level Classification model with training-validation sets that have different sizes. In the research conducted, it can be concluded that the RoBERTa algorithm with training-validation 4065 users has the highest f1-score value of 78, 0 and the XLNet algorithm with training-validation sets of 4065 users has the highest accuracy value of 77, 1 [10].

Furthermore, Xiaofeng Wang, et al. conducted research that explored the potential of deep learning methods in predicting the risk of depression from Chinese Microblogs. The deep learning methods used are BERT, RoBERTa, and XLNet. From the results of the study, the RoBERTa and XLNet methods showed great superiority. RoBERTa achieved the best performance with an F1-Score value of 0.422 at the risk of level 1 depression (mild depression) and XLNet obtained an F1-Score value of 0.493 at the risk of level 2 depression (moderate depression) and an F1-Score value of 0.445 at the risk of level 3 depression (severe depression) [11]. The analysis that has been done, it shows some errors in predicting the level of depression because there are many ambiguous words in Chinese Microblogs that are difficult to distinguish independently [11].

This research focuses on depression detection on twitter social media using the XLNet method which introduces Permutation Language Modeling [12]. The XLNet model is built using extra-long transformers so that it can capture longer dependencies from sentences efficiently [13]. In addition, this research also utilizes the DASS-42 (Depression Anxiety Stress Scales) in the form of a questionnaire filled out by respondents related to self-report consisting of 42 negative emotional symptoms needed to measure risk group assessment [14], especially the depression labeling of the dataset that will be used during the research.

II. Research Methodology

The system built in this research can detect a person's depression through twitter social media which is seen based on user tweets using the XLNet method. In detecting depression, there are several stages, namely data collecting, data preprocessing, system modeling, and evaluation. An overview of the system built during the research can be seen in Figure 1.

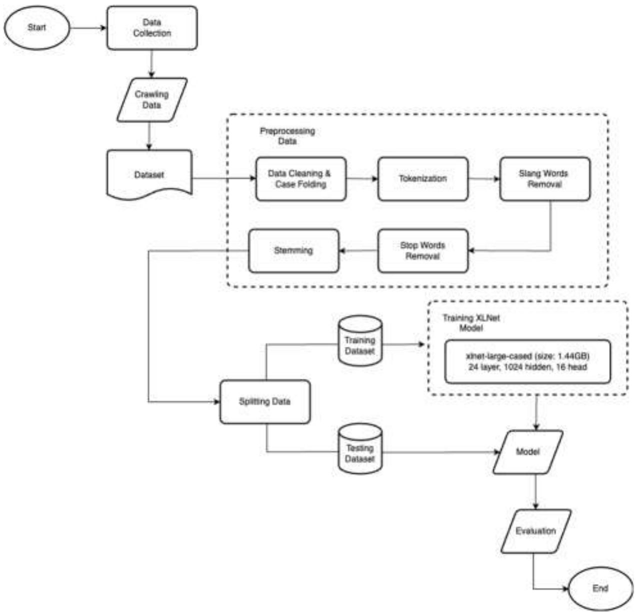

Figure 1. Flowchart System

A. Data Collection

Data collection is the first stage carried out in research. Data collection was carried out by distributing DASS-42 questionnaires which were filled in by respondents. Depression Anxiety Stress Scales (DASS-42) is a self-report of depression, anxiety, and stress developed by Lovibond & Lovibond [14]. The DASS-42 is a self-report questionnaire consisting of 42 negative emotional symptoms from the depression, anxiety, and stress scales. Respondents will fill out a questionnaire that has 14 items on each of the three scales using a 0-3 scale according to the extent of the condition that the respondent has experienced on each symptom contained in the questionnaire [14]. The DASS-42 score can be seen in Table I [15]

TABLE I
DASS-42 (DEPRESSION ANXIETY STRESS SCALES) SCORES

Level of	Depression	Anxiety	Stress
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Medium	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	>28	>20	>34

TABLE II
DASS-42 FACTOR LOADING ITEMS

Scale	DASS-42 Factor Loading Items
Depression	3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42
Anxiety	2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41
Stress	1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39

The research only focuses on detecting depression so that the depression labeling process on respondents only calculates the number of depression scale scores obtained by respondents with several items on DASS-42. Some of the factor loading items on the DASS-42 can be seen in Table II [16].

Furthermore, crawling data is obtained from the twitter tweets of respondents who have filled out the DASS-42 questionnaire. The results of crawling data are then labeled based on the calculation of the DASS-42 questionnaire score which is labeled 1 if depression is indicated with a total score obtained by respondents of more than 9 and labeled 0 if depression is not indicated with a total score obtained by respondents between 0-9. The following are the results of crawling data obtained from 159 datasets with labeling 1 indicating depression in as many 94 data and labeling 0 not indicating depression in as many 65 data can be seen in Figure 2.

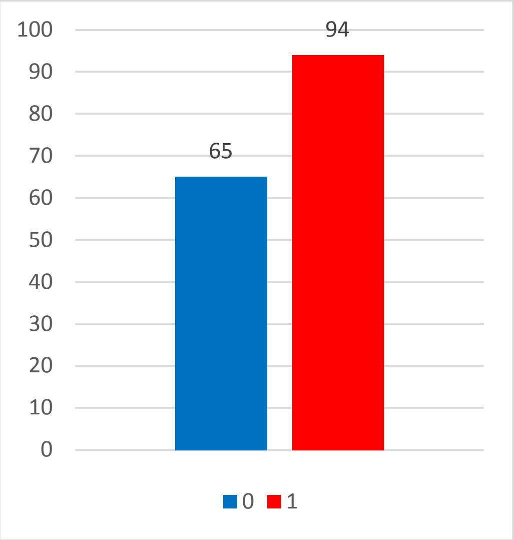

Figure 2. Comparison of Number of Datasets with Labels 0 and 1

Sample data used during the study that has been calculated based on DASS-42 scores consisting of usernames, tweets, and labels can be seen in Table III.

TABLE III
SAMPLE DATASET

User	Tweet	Label
Username 1	@polbanfess Ya beda-beda atuh, ga bisa disamaratakan @polbanfess Coba minta daftar tempat yg pernah di-KP-in ke angakatan-angkatan sebelumnya Awkward banget untuk introvert sepertiku @polbanfess Dulu kalo kebetulan cuma ada uang 50/100 selembar,	1
Username 2	ya allah mau punya cewe muka bad bitch:((gagitu konsepnyaa, we still can change this. hold on plisss @pengencptkaya Yg di tiktok bukan? @minorneeds	0
Username 3	@telyufess Org sukabumi pada tilil gini kenapa ya? @dezanahmd @idextratime coba mnh nyermin jan bari cosplay suster si mengmong udah meninggal sekarang Nempel sticker di kulkas jaman dulu satispaying. RIP 2002-2022	1

B. *Preprocessing Data*

The dataset that has been collected will then enter the data preprocessing stage first. Data preprocessing is a process that changes the form of data that has been collected into a form of data that is easy to understand. Data preprocessing is important to be able to simplify the data processing process that will be used. The data preprocessing carried out during the research consists of data cleaning suvch as removing username, removing retweet text 'RT', removing hyperlinks, removing punctuations, removing hashtags, removing commas, removing numbers, removing whitespaces, removing new lines, and case folding which is changing the entire text contained in the data into lowercase letters [17], tokenization is a method that separates a sentence into per-word [18], slang words removal is a method to change a word that is not formal/unstandard into a formal word, stop words removal is a method that removes a common word that is often used such as the word "is", "it", "this", and others so that the selected word is a word that is considered important only [17], and stemming is a method that describes the form of a word into its basic word [17]. The results of data preprocessing carried out during the study can be seen in Table IV.

TABLE IV
PREPROCESSING DATA

Preprocessing Method	Tweet
Raw Dataset	DEPRESI sungguh mengganggu [URL], slh satu cara mengatasi depresi adlh berkonsultasi, [@username] aku mengalami sedikit depresi!!!!!
Data Cleaning & Case Folding	depresi sungguh mengganggu slh satu cara mengatasi depresi adlh berkonsultasi aku mengalami sedikit depresi
Tokenization	“depresi”, “sungguh”, “mengganggu”, “slh”, “satu”, “cara”, “mengatasi”, “depresi”, “adlh”, “berkonsultasi”, “aku”, “mengala-mi”, “sedikit”, “depresi”
Slang Words Removal	“depresi”, “sungguh”, “mengganggu”, “salah”, “satu”, “cara”, “mengatasi”, “depresi”, “adalah”, “berkonsultasi”, “aku”, “mengalami”, “sedikit”, “depresi”
Stop Words Removal	“depresi”, “sungguh”, “mengganggu”, “salah”, “satu”, “cara”, “mengatasi”, “depresi”, “berkonsultasi”, “aku”, “mengalami”, “sedikit”, “depresi”
Stemming	“depresi”, “sungguh”, “ganggu”, “salah”, “satu”, “cara”, “atasi”, “depresi”, “konsultasi”, “aku”, “alami”, “sedikit”, “depresi”

C. *Splitting Data*

After preprocessing the data on the dataset used, then enter the data splitting stage. At the data splitting stage, the dataset is divided into two parts, namely training data as training data in building the model and testing data as test data to determine the performance of the model building. In the research, testing scenarios were carried out with different splitting data train sizes and test sizes to test which data division could produce the best accuracy value during the research.

D. *Pretraining Model XLNet*

XLNet is an autoregressive (AR) language modeling that learns bidirectional context. XLNet introduces Permutation Language Modeling (PLM) which maximizes the expected log probability based on the order of factorization of all possible sequences [12]. XLNet can learn bidirectional contexts because there is a permutation operation consisting of right and left tokens and it can read information from all positions so that XLNet can effectively overcome the independence assumption [19].

XLNet has a two-stream self-attention architecture. Content stream attention is like standard self-attention that can access content information, while query stream attention does not have access to content-related information [19]. In XLNet, permutations can be obtained with mask-attention but since query stream attention cannot access information, it can therefore be overcome with the two-stream self-attention method that can present target-aware. Figure 3 shows the architecture of XLNet [19].

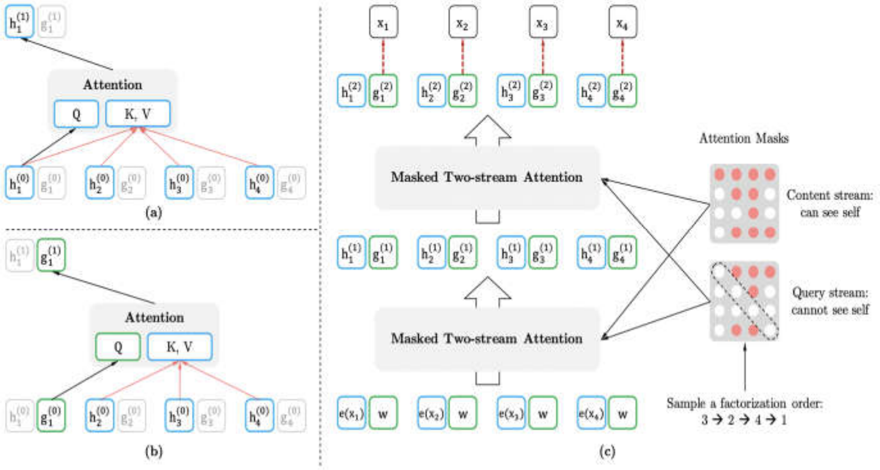
Attention
Q
K, V
(a)
Attention
Q
K, V
(b)
Masked Two-stream Attention
Masked Two-stream Attention
Attention Masks
Content stream: can see self
Query stream: cannot see self
Sample a factorization order: 3 → 2 → 4 → 1
(c)

Figure 3. Architecture of XLNet

The following are the equations (1) and (2) used in presenting the target-awareness with the two-stream self-attention method.

$$g_{z_t}^{(m)} \leftarrow \text{Attention}\ (\text{Q} = g_{z_t}^{(m-1)}, \text{KV} = h_{z_{<t}}^{(m-1)};\ \theta), \qquad (1)$$

$$h_{z_t}^{(m)} \leftarrow \text{Attention}\ (\text{Q} = h_{z_t}^{(m-1)}, \text{KV} = h_{z_{\leq t}}^{(m-1)};\ \theta), \qquad (2)$$

Description:

$h_\theta(x_{z_{\leq t}})$: representation of the content that can access the context x_{z_t} itself.

$g_\theta(x_{z_{<t}}, z_t)$: representation of a query that can only access the contextual information of $x_{z_{<t}}$ and position z_t but cannot access the content of x_{z_t}.

E. Evaluation

The evaluation process during the research was carried out using a confusion matrix. The confusion matrix is a concept that contains information related to actual classification and prediction. The confusion matrix has two dimensions, one dimension contains the class of an object, and the other dimension contains other classes to be classified [20]. In the confusion matrix, there are 4 categories in presenting the data owned, including [21]:

1. True Positive (TP) i.e., the predicted data is positive, and the actual data is positive.
2. False Positive (FP), which is data that is predicted to be positive, and the actual data is negative.
3. True Negative (TN), that is, the predicted data is negative, and the actual data is positive.
4. False Negative (FN), which is data that is predicted to be negative, and the actual data is negative.

Based on the four categories, it will be used to evaluate the model. The evaluations that will be carried out are accuracy, precision, recall, and f1-score which can be calculated using formulas (3), (4), (5), and (6).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\text{-}Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (6)$$

III. RESULTS AND DISCUSSION

The test results that have been carried out in the study can be seen from several test scenarios that are made to analyze the results of the accuracy performance of a person's depression detection using the XLNet method. The test scenarios carried out include:

A. Test Scenario 1

Test scenario 1 is done by applying oversampling which is one of the random sampling techniques that is usually used to balance the amount of data in two classes by increasing the number of minority class samples in the data by repeating several examples in the sample data [22]. In this research, the oversampling technique is used because it has a small amount of data. Oversampling can help to get the expected training value. Furthermore, pre-processing is carried out using Stemming and without Stemming to see whether the Stemming preprocessing stage affects the resulting accuracy value. In addition, data division is carried out on different train sizes and test sizes, namely: 90:10, 80:20, 70:30, and 60:40 to find out which data division can achieve the best accuracy value. Another test is to test the tuning hyperparameters contained in the XLNet model, including the epoch value which is a cycle that is run on the neural network during data training. The epoch value tested is different, namely epoch = 5,15, 25. Another tuning hyperparameter is batch size testing which is the number of data samples used for each epoch spread on the neural network during the training process [23]. The number of batch sizes tested, namely, batch size = 8, 16. The overall results of test 1 can be seen in Table V.

The results of test scenario 1 show that research using the data preprocessing stage without using stemming and train size 90% & test size 10% as well as epoch = 25, and batch size = 16 achieved the highest accuracy value, which is 95%. Based on the results of test scenario 1, it can be concluded that the data preprocessing stage using stemming gets lower accuracy results compared to the preprocessing stage without stemming because stemming at the preprocessing stage will change a word form into its basic word so that it can affect the meaning of a sentence. The meaning of a sentence will change into a different meaning. Therefore, the data preprocessing stage without using stemming is better for use in research. In addition, from the table of test scenario results 1, the larger the epoch value does not always have a better accuracy value compared to the smaller epoch value even though it uses a smaller batch size or a larger batch size value.

TABLE V
TEST SCENARIO 1 RESULTS

Preprocessing Data	Split Dataset	Batch Size	Epoch	Accuracy
Stemming	90:10	8	5	55%
			15	80%
			25	50%
		16	5	55%
			15	65%
			25	70%
	80:20	8	5	48%
			15	48%
			25	73%
		16	5	48%
			15	80%
			25	80%
	70:30	8	5	50%
			15	82%
			25	85%
		16	5	48%
			15	48%
			25	75%
	60:40	8	5	50%
			15	50%
			25	50%
		16	5	51%
			15	70%
			25	66%
No Stemming	90:10	8	5	55%
			15	85%
			25	65%
		16	5	50%
			15	70%
			25	**95%**
	80:20	8	5	57%
			15	78%
			25	70%
		16	5	53%
			15	55%
			25	48%
	70:30	8	5	53%
			15	48%
			25	70%
		16	5	48%
			15	73%
			25	87%
	60:40	8	5	54%
			15	57%
			25	51%
		16	5	59%
			15	53%
			25	81%

B. *Test Scenario 2*

Test scenario 2 is carried out using the data preprocessing stage without stemming with epoch = 25 and batch size = 16 and the division of train size 90% and test size 10% because it gets a very good accuracy value performance in the previous test. Furthermore, test scenario 2 will be tested with different input size values to find out whether the amount of input size can affect the accuracy level during the research conducted. The results of test scenario 2 show that the input size value = 64 gets the highest accuracy value of 95%. Based on the results of test 2, it can be concluded that the more the input size value, the better the accuracy value. The results of test 2 can be seen in Table VI.

TABLE VI
TEST SCENARIO 2 RESULTS

Preprocessing Data	Split Dataset	Batch Size	Epoch	Input Size	Accuracy
No Stemming	90:10	16	25	8	80%
				16	80%
				24	85%
				32	90%
				64	**95%**

C. *Test Scenario 3*

Test scenario 3 was conducted with additional hyperparameter tuning, namely input size with a value of 64 because the previous test achieved the best accuracy value. Another test is to compare different learning rate values on the XLNet model. Learning rate is several weights that will continue to be updated during the training process to minimize the loss of function value on the neural network. The results of test scenario 3 show that the learning rate with the smallest value, namely, 1e-6 has the best accuracy value of 90%. Based on this, it can be concluded that the smaller the learning rate value, the better the accuracy value because it affects the training process. If the learning rate value is small, the training process will be longer, but the accuracy of the model built can increase with a slow training process. Conversely, if the learning rate value is large, the accuracy of the model built can decrease because it has a faster training process. The results of test 3 can be seen in Table VII.

TABLE VII
TEST SCENARIO 3 RESULTS

Input Size	Learning Rate	Accuracy
64	1e-3	50%
	1e-4	50%
	1e-5	75%
	1e-6	**90%**

The research conducted focuses on how to build a model to achieve the best accuracy value by testing different values of hyperparameter tuning of the XLNet model which is divided into 3 test scenarios. The test results can be concluded that the model built achieves the best performance with a value of 95% in test scenario 1 by applying oversampling, without stemming at the data preprocessing stage with a train size of 90% & test size 10% and epoch value = 25 and batch size = 16. In test scenario 2, testing on one of the tuning hyperparameters with an input size value of = 64 gets a good accuracy value of 95%. In test scenario 3, testing one of the tuning hyperparameters with a learning rate value = 1e-6 achieved a good accuracy value of 90%. This shows that the XLNet model built for depression detection on twitter social media achieves a good performance value with an average accuracy value of 93.33% based on 3 test scenarios that have been carried out. Information on the average value of accuracy obtained for each test scenario that has been carried out during the research can be seen in Table VIII.

TABLE VIII
DESCRIPTION OF THE AVERAGE VALUE OF ACCURACY

Scenario Test	Accuracy
Test 1	95%
Test 2	95%
Test 3	90%
Average	**93.33%**

There is research related to the identification of optimism and pessimism through messages on twitter using the XLNet model and the Deep Consensus algorithm conducted by Ali Alsharani, et al. The research used XL-Net-Base and XLNet-Large by dividing 80% for training, 10% for evaluation, and 10% for testing. The research used hyperparameters with different values including learning rate = 2e-4, 2e-5, and 2e-6, batch size = 16, 32, 64, maximum input length = 32, 64, 128 and selected the most stable and efficient value for further testing. The fine-tuning procedure utilizes the representation with CLS which is a special token for classification. The XLNet-Large model achieves an accuracy value of 96.45% for level 1 tweets and 85.28% for level 0 tweets. XLNet-Large gets a minimum accuracy value of 95.83% and a maximum accuracy value of 97.92% on level 1 tweets. While the XLNet-Base model gets a minimum accuracy value of 95.57% and a maximum accuracy value of 97.39% at tweet level 1. The evaluation results on the research model built to provide the best accuracy on the benchmark dataset [13].

Another research was conducted by Gabriel Raka, et al. related to how to filter, and categorize research articles on COVID-19 that apply a text classification system to make it easier for doctors to save time by pre-filtering labeling articles for EBM. The research conducted compared the results of document classification with the Random Forest method with a special tokenizer created by Epistemonikos, the XLNet model which uses a linear layer as a classification, and the BioBERT language model. The results showed that the XLNet model achieved the best performance by getting the highest F1 score for each category of document types such as the sys-thematic review achieving an F1 value of 0.97 compared to the value of other models. In this case, XLNet is more flexible because it can sufficiently train embeddings and classification processes regardless of document categories [24].

From the research conducted related to depression detection on twitter social media using the XLNet method and several other studies, especially related to text classification which also uses the XLNet method, each of the studies can be concluded to achieve the best accuracy value. This is because XLNet is a general autoregressive model that reaches SOTA (state-of-the-art) in many NLP (Natural Language Processing) tasks [13]. In addition, XLNet learns bidirectional context so that it can effectively overcome the independence assumption [19].

IV. CONCLUSION

Based on 3 test scenarios that have been carried out in the research, the built model achieves the best performance by obtaining an average accuracy value of 93.33%. In test scenario 1 applying over-sampling, without stemming at the data preprocessing stage with train size 90% & test size 10% and epoch value = 25, and batch size = 16 produces a good accuracy value of 95%. However, a large epoch value does not prove to have a better accuracy value compared to a smaller epoch value despite using a smaller batch size or a larger batch size value. In test scenario 2 using one of the tuning hyperparameters with an input size value of = 64, a good accuracy value of 95% was obtained. This shows that the input size value can affect the accuracy value of the model. The greater the input size value, the better the accuracy value. In test scenario 3 using one of the tuning hyperparameters with a learning rate value = 1e-6 achieved a good accuracy value of 90%. This is because the learning rate affects the training process. The training process will be longer because it uses a small learning rate value but the accuracy during the training process of the built model can increase. In future research, more datasets can be tried with the use of feature extraction to get better performance results and try different values on one of the XLNet tuning hyperparameters, namely batch size to find out the effect of batch size during the research conducted.

REFERENCES

[1] World Health Organization (WHO), "World Health Organization (WHO) : Depression," 13 September 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed 28 October 2021].

[2] K. Ho, "YouGov : A quarter of Indonesians have experienced suicidal thoughts," 26 June 2019. [Online]. Available: https://id.yougov.com/en-id/news/2019/06/26/quarter-indonesians-have-experienced-suicidal-thou/. [Accessed 28 October 2021].

[3] D. Ridge and S. Ziebland, "'The old me could never have done that': How people give meaning to recovery following depression," Qualitative Health Research, vol. 16, no. 8, pp. 1038–1053, Oct. 2006, doi: 10.1177/1049732306292132.

[4] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," Current Opinion in Behavioral Sciences, vol. 18. Elsevier Ltd, pp. 43–49, Dec. 01, 2017. doi: 10.1016/j.cobeha.2017.07.005.

[5] S. KEMP, "DIGITAL 2022: GLOBAL OVERVIEW REPORT," Data Reportal, 26 January 2022. [Online]. Available: https://datareportal.com/reports/digital-2022-global-overview-report. [Accessed 26 August 2022].

[6] "SPECIAL REPORT DIGITAL 2021," We Are Social, January 2021. [Online]. Available: https://wearesocial.com/digital-2021. [Accessed 26 August 2022].

[7] Statista, "Statista : Leading countries based on number of Twitter users as of October 2021," October 2021. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed 28 October 2021].

[8] A. Priya, S. Garg, and N. P. Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms," in Procedia Computer Science, 2020, vol. 167, pp. 1258–1267. doi: 10.1016/j.procs.2020.03.442.

[9] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep Learning for Depression Detection of Twitter Users," 2018.

[10] Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, and J. Luo, "Monitoring Depression Trend on Twitter during the COVID-19 Pandemic," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.00228

[11] X. Wang et al., "Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis," JMIR Medical Informatics, vol. 8, no. 7, Jul. 2020, doi: 10.2196/17958.

[12] Y. Wang, J. Zheng, Q. Li, C. Wang, H. Zhang, and J. Gong, "Xlnet-caps: Personality classification from textual posts," Electron., vol. 10, no. 11, pp. 1–16, 2021, doi: 10.3390/electronics10111360.

[13] A. Alshahrani, M. Ghaffari, K. Amirizirtol, and X. Liu, "Identifying Optimism and Pessimism in Twitter Messages Using XLNet and Deep Consensus," Proc. Int. Jt. Conf. Neural Networks, 2020, doi: 10.1109/IJCNN48605.2020.9206948.

[14] N. Bilgel and N. Bayram, "Depresyon anksiyete stres ölçeğinin (DASS-42) Türkçeye uyarlanmiş şeklinin psikometrik özellikleri," Noropsikiyatri Arsivi, vol. 47, no. 2, pp. 118–126, 2010, doi: 10.4274/npa.5344

[15] N. Syafitri, Y. Arta, A. Siswanto, and S. P. Rizki, "Expert System to Detect Early Depression in Adolescents using DASS 42," Oct. 2020, pp. 211–218. doi: 10.5220/0009158202110218.

[16] M. M. Antony, B. J. Cox, M. W. Enns, P. J. Bieling, and R. P. Swinson, "Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample," Psychol. Assess., vol. 10, no. 2, pp. 176–181, 1998, doi: 10.1037/1040-3590.10.2.176.

[17] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 4, no. 3, pp. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

[18] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," 2019 Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.

[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. v. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.08237

[20] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," Information Sciences, vol. 340–341, pp. 250–261, May 2016, doi: 10.1016/j.ins.2016.01.033.

[21] S. Wagner, Association for Computing Machinery, and ACM Digital Library., PROMISE : 8th International Conference on Predictive Models in Software Engineering : Lund, Sweden, Sept 21-22, 2012 : co-located with ESEM 2012.

[22] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, no. May, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[23] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," ICT Express, vol. 6, no. 4, pp. 312–315, 2020, doi: 10.1016/j.icte.2020.04.010.