Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 1, February 2024, pp. 121-128

DOI: https://doi.org/10.52436/1.jutif.2024.5.1.1187
p-ISSN: 2723-3863
e-ISSN: 2723-3871

DEPRESSION DETECTION ON TWITTER USING GATED RECURRENT UNIT

Alfransis Perugia Bennybeng Holle[1], Warih Maharani[2]

[1,2]Informatics, Informatics Faculty, Universitas Telkom, Indonesia, Indonesia
Email: [1]alfransholle@student.telkomuniversity.ac.id, [2]wmaharani@telkomuniversity.ac.id

(Article received: July 04, 2023; Revision: July 22, 2023; published: January 31, 2024)

Abstract

In the present era, technological advancements have significantly impacted society, particularly in the use of social media. One popular social media platform is Twitter, where people could share moments, thoughts, and statuses. However, since the COVID-19 pandemic, the usage of Twitter increased, and some users began exhibiting symptoms of depression. The condition of depression required a means to channel emotions that could assist users in coping. By employing the GRU method and Word2Vec feature extraction, we developed a depression detection system capable of analyzing users' Twitter posts and identifying potential signs of depression. The dataset used in this research was obtained from 165 participants who agreed to utilize their personal Twitter data and completed a questionnaire based on the Depression Anxiety and Stress Scales-42 (DASS-42). The questionnaire results served as labels that were processed for Word2Vec feature extraction and subsequently fed into the GRU model. The evaluation revealed an accuracy rate of 57.58% and an f1-score of 56.25. By using the bidirectional layer in the model, there is an improvement in precision, recall, and f1-score values.

Keywords: *Dataset, Depression, Evaluation, GRU, Twitter.*

1. INTRODUCTION

In the era of advancing technology like today, technological developments have reached an incredible level and have had a significant impact on society. One of the main impacts of this progress is the use of social media. Through platforms such as TikTok, Instagram, Facebook, and Twitter, individuals have the opportunity to express their true selves more widely. Social media has become a popular space for people to interact and share content with others.

One of the highly popular social media platforms is Twitter, which was launched in 2006. With Twitter, users can easily send and read short posts or tweets, making it an ideal platform for sharing information and communicating succinctly. In fact, during the COVID-19 pandemic, Twitter usage increased by 34% in the second quarter of 2020 [1]. According to data from dataindonesia.id, Indonesia is among the countries with the highest number of Twitter users globally, with a staggering 18.45 million users as of 2022 [2]. This indicates that social media has become an important part of many people's daily lives, especially during challenging times like a pandemic.

However, during the COVID-19 pandemic, there has also been a significant increase in the number of individuals experiencing depression. It has been reported that there was a 25% increase in the number of individuals experiencing depression [3]. Based on the results of the Indonesia National Adolescent Mental Health Survey (I-NAMHS), approximately 15.5 million adolescents aged 10-17 years in Indonesia experienced mental health issues in 2022 [4]. In a study conducted by Andria Pragholapat [5], the COVID-19 pandemic has had an impact on students' mental health. Depression is a mental disorder commonly characterized by loss of interest, persistent feelings of guilt, fatigue, and difficulty concentrating. This condition can persist for a long and recurring period [6]. In facing this depressive condition, individuals greatly need means to channel their emotions and seek support.

In this context, research is conducted to develop a method using Gated Recurrent Unit (GRU) in analyzing users' Twitter posts or tweets to detect depression. GRU is a type of recurrent neural network (RNN) architecture that has the ability to understand previous contexts and has advantages in addressing sequence processing problems. The input and output format of a GRU resembles that of a regular RNN), but its internal architecture is more akin to that of a Long Short-Term Memory (LSTM) network [7]. In the analysis stage, users' tweets will be categorized into two categories: positive and negative, based on the frequently used words by individuals experiencing depression. Some examples of words that fall into the positive category are "happy," "beach," and "photo," while the negative category includes words like "death," "no," and "never" [8]. Through this analysis, it is expected to identify patterns of words related to depression, enabling more accurate detection.

Previously, the use of GRU method has been proven successful in research on detecting hate speech by Junanda Patihullah [9], achieving an

accuracy of 92.96%, and detecting hand gestures in videos by Gulpi Qorik Oktagalu Pratamasunu [10], achieving an accuracy of 88.00%. Therefore, this method shows strong potential for use in detecting depression through tweet analysis. Several previous studies have been conducted to detect depression, such as the one conducted by Putri Simolang [11], using the BiLSTM (Bidirectional Long-Short Term Memory) method in a similar analysis. However, this research only achieved an accuracy rate of 53.12% and an f1-score of 68.08%. Another study conducted by Hafshah Haudli Windjatika [12] used the LSTM (Long-Short Term Memory) method, achieving an accuracy rate of 77.95% and an f1-score of 51.14%. The subsequent research was conducted by Andre Agasi Simanungkalit [13], where the BilLSTM (Bidirectional Long-Short Term Memory) method was employed. The findings of the study revealed that the method achieved an accuracy of 70.59% and an f1-score of 80%.

In this research, the GRU method will use word2vec as a feature extraction technique to enhance accuracy and f1-score in detecting depression through tweet analysis. Word2vec is a natural language processing technique used to represent words in the form of numerical vectors [14]. This method generates vector representations based on the contextual words found in the given text corpus. By using word2vec, words that have contextual similarities will have vector representations that are closer to each other in the vector space. This allows the GRU model to obtain a better understanding of the relationships between words in tweet texts, which can improve the model's performance in detecting depression. Twitter user data collected using the Twitter API will be cleaned and processed to be used as training data to train the GRU model. Subsequently, the trained GRU model will be tested on unseen tweet data to evaluate its performance.

As part of the research, Twitter users will be asked to fill out the DASS-42 (Depression Anxiety and Stress Scales) questionnaire consisting of 42 symptoms[15]. The main purpose of using the DASS-42 questionnaire is to provide labels to the collected dataset for better analysis.The goal of this research is to identify depression on the Twitter social media platform by detecting and analyzing tweets and providing recommendations if the method used successfully processes the limited amount of data. By using this limited data, we can determine the model's success rate by examining its accuracy.

This research primarily focuses on the detection of tweets in the Indonesian language, but for other languages such as English and local languages, they will be translated using a prepared dictionary. The results of this research can be used by companies, especially in Indonesia, in the employee recruitment process. Thus, the research findings can serve as a benchmark to determine whether Twitter users are experiencing depression or not based on their Twitter usage.

2. RESEARCH METHOD

The objective of this study was to develop a system that could identify depression using a Gated Recurrent Unit (GRU). The data utilized was collected from the Twitter social media platform. In order to carry out this detection process, the system underwent several steps, such as data crawling, data preprocessing, data splitting, model training, and evaluation. The process was illustrated in Figure 1.

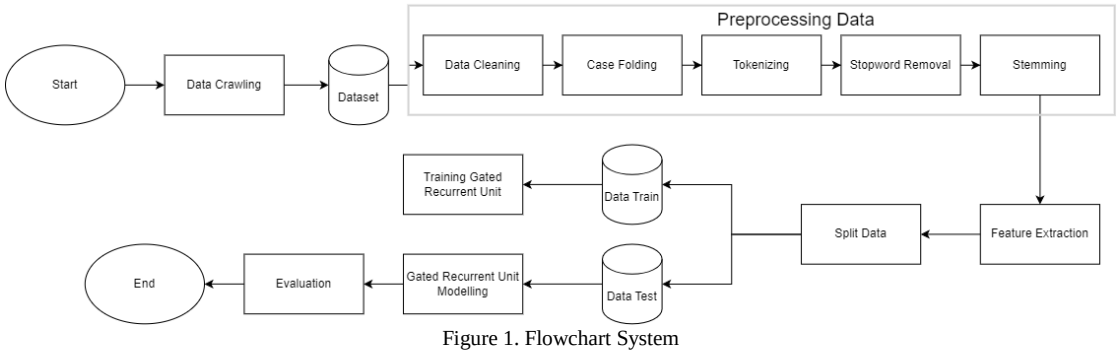

Figure 1. Flowchart System

2.1. Crawling Data

The data used in this research consisted of respondent tweets and questionnaire results. The tweet data was obtained using the Twitter API, which served as a connection between the developed system and Twitter. Meanwhile, the questionnaire used was the DASS-42 questionnaire, designed to measure negative emotional states such as depression, anxiety, and stress. The results from the DASS-42 questionnaire were used to assign the positive or negative labels to the data. After obtaining both datasets, the data were merged and inserted into a single CSV file. According to Table 1, in this study, respondents who scored 28 or above were classified as severely depressed, indicating a extreme level of severity.

Table 1. The Severity of The Disorder

Disorder	**Severity Level**				
	Normal	Low	Medium	Heavy	Extreme
Depression	0-9	10-13	14-20	21-27	28+

2.2. Pre-processing Data

During the data preprocessing stage, a series of actions were undertaken to ready the data for model processing. This involved enhancing the data quality and converting it into a format that the model could comprehend. By completing this stage, the data was prepared for the subsequent phase. The steps involved in data preprocessing were as follows:

2.2.1. Data Cleaning

In the first step, the text will be cleaned by removing numbers, punctuation marks, symbols, non-alphabet characters in the Indonesian language, excessive spaces, and converting abbreviations into their full words.

2.2.2. Case Folding

In the subsequent stage, known as case folding, the text undergoes a transformation to convert any capital letters into lowercase. This is done to ensure uniformity and consistency in the text. By converting all letters to lowercase, the text can be standardized, making it easier for further processing and analysis.

2.2.3. Tokenizing

Tokenization played a crucial role in separating sentences into a sequence of words. This process was essential for further analysis and deeper understanding of text data. By breaking down the text into tokens, each word could be treated as an individual entity, enabling various language processing tasks such as sentiment analysis, topic modeling, and text classification.

2.2.4. Stopword Removal

This process involves removing words that have no meaning or are commonly referred to as stop words, as well as eliminating duplicate words in the text. The purpose of this process is to enhance the quality of the text and eliminate irrelevant elements for analysis.

2.2.5. Stemming

This process involves removing affixes from words to obtain their base form. By eliminating affixes, we can derive the root or base word that represents the core meaning. This process helps reduce the dimensionality of words and optimize text analysis.

2.3. Feature Extraction

Feature extraction is the process in which data is represented by structured values. The feature extraction method used in this research is word2vec. Word2vec is an algorithm that connects each word in the text with a vector[16]. This algorithm was developed by Mikolov and his team in 2013 and has been widely used in NLP research. Word2vec represents words as vectors that store semantic information, using unsupervised learning models with neural networks consisting of hidden layers and fully connected layers. The weight matrix in the hidden layer is used to transform words into vectors. Word2vec relies on local language information, where the semantics of words are learned based on the words around them. The word2vec algorithm consists of two types, Continuous Bag-of-Word (CBOW) and Skip-gram[17]. CBOW uses context to predict the target word, while Skip-gram uses a word to predict its context.

2.4. Training GRU Model

The Gated Recurrent Unit (GRU) was an architecture created by Kyunghun Cho in 2014. The goal of GRU was to make each recurrent unit capable of capturing each relationship (dependency) at different times[10]. The GRU has been motivated by the LSTM unit[18]. GRU had 2 gates, namely the reset gate and the update gate. The reset gate functioned to regulate how much information from the past would be forgotten or ignored by the GRU unit, while the update gate functioned to control how far new information would be integrated into the GRU unit.

2.5. Confusion Matrix

The Confusion Matrix is a evaluation method used to measure the performance of a model. It consists of four values: True Positive (TP) for correct positive predictions, True Negative (TN) for correct negative predictions, False Positive (FP) for incorrect positive predictions, and False Negative (FN) for incorrect negative predictions. By using the Confusion Matrix, we can calculate evaluation metrics such as accuracy, precision, recall, and so on to assess the prediction quality of the model. Here is table 2 that represents the confusion matrix.

Table 2. Confusion Matrix

Actual Class	**Predicted Class**	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The confusion matrix was used as an evaluation method in machine learning algorithms [19], where the four values could measure accuracy, precision, recall, and F1-score. Accuracy was a parameter that measured the percentage of cases correctly predicted out of the total cases. Precision was the ratio of correctly predicted positive cases to the total predicted positive cases. Recall was the ratio of correctly predicted positive cases to the total observations in the actual class. F1-score was a measure used in situations where the class distribution was imbalanced, especially if there were a large number of true negative observations. The

formulas for these four metrics could be found in equations 1, 2, 3, and 4 [20].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

3. RESULT

A dataset was collected from 165 users, with approximately 100 tweets obtained from each user. The collected data was labeled using the DASS-42 questionnaire. The distribution of the labeled data could be observed in table 3, revealing that the data generated was imbalanced, with an uneven distribution among the labels.

Table 3. Data Distribution

Data Train		Data Test	
Positive	Negative	Positive	Negative
82	50	17	16

3.1. Preprocess Data Result

The data will undergo preprocessing before entering the classification process. The preprocessing steps include data cleaning, case folding, tokenizing, stopword removal, and stemming. The preprocessed data can be seen in table 4.

Table 4. Preprocessing Data Result

Process	Input	Output
Data cleaning	@thestorykids @StrayKids Baik banget dong our artistik man Punten kakak kaka bisa bantu isi kuisioner ini gak buat tugas. Terima kasih https://t.co/vooj98riMw	Baik banget dong our artistik man Punten kakak kaka bisa bantu isi kuisioner ini gak buat tugas Terima kasih
Case folding	Baik banget dong our artistik man Punten kakak kaka bisa bantu isi kuisioner ini gak buat tugas Terima kasih	baik banget dong our artistik man punten kakak kaka bisa bantu isi kuisioner ini gak buat tugas terima kasih
Tokenizing	ya btul tp mls dikira oversharing sini cerita sm aku lagu lagunya	['ya', 'btul', 'tp', 'mls', 'dikira', 'oversharing', 'sini', 'cerita', 'sm', 'aku', 'lagu', 'lagunya']
Stopword removal	['ya', 'btul', 'tp', 'mls', 'dikira', 'oversharing', 'sini', 'cerita', 'sm', 'aku', 'lagu', 'lagunya']	['malas', 'oversharing', 'cerita', 'lagu', 'lagunya']
Stemming	['malas', 'oversharing', 'cerita', 'lagu', 'lagunya']	['malas', 'oversharing', 'cerita', 'lagu']

3.2. Feature Extraction Result

In this study, word2vec is used as the feature extraction method. The word2vec settings include a window size of 5, vector size of 300, min_count of 10, and 4 workers. The extraction results for the words "putus" and "hati" are transformed into vectors of length 300. The extraction results can be observed in Table 5.

Table 5. Word2Vec Result

Words	Word2Vec Vector
putus	6.46811211e-03, 3.05115670e-01, … , -9.14795995e-02
hati	-8.52851104e-03, 3.51242006e-01, … , -8.00894573e-02

Using word2vec, we can explore the most similar words, where word2vec detects the similarity between words we are searching for. The table 6 presents the list of the most similar words that can be observed.

Table 6. Most Similar Words

Words	Most Similar Words
putus	[('pacar', 0.9922822713851929), ('bagus', 0.9902629852294922), ('dingin', 0.9899163246154785), ('orang', 0.9898964166641235), ('ajak', 0.9897858500480652), ('kali', 0.9896864891052246), ('parah', 0.9886367321014404), ('kesal', 0.9871484041213989), ('asli', 0.9869505167007446), ('peras', 0.9863447546958923)]
hati	[('lelah', 0.9908594489097595), ('malas', 0.9903266429901123), ('minggu', 0.9902850985527039), ('hidup', 0.9900622367858887), ('jaga', 0.989186704158783), ('seru', 0.9888458847999573), ('makan', 0.9884860515594482), ('rasa', 0.9881348609924316), ('pusing', 0.9877614378929138), ('kepala', 0.9876381158828735)]

3.3. Test Analysis

In this stage, experiments were conducted and data was analyzed to find the best model using hyperparameter tuning. This stage was divided into two scenarios. The first scenario incorporated a bidirectional layer, while the second scenario did not. The obtained results were analyzed and presented as the outcomes of these experiments.

3.3.1. Test Analysis on Gated Recurrent

In this scenario, the model was tested by dividing the data into different ratios for each model: 90:10, 80:20, and 70:30. Using the imbalanced data, the testing yielded the best results with a ratio of 70:30, achieving an accuracy of 56.00%, precision of 56.00%, recall of 100%, and an f1-score of 71.79%.

Due to the recall value of 100% and the accuracy being equal to the precision, the model performed best in identifying and classifying positive and negative cases as non-existent. The results of each testing can be seen in the following table 7.

Table 7. Imbalance Data without Layer Bidirectional

Ratio	Batch Size	Data	Accuracy	Precision	Recall	F1-Score
70:30	16	Train	60,87%	60,87%	100,00%	75,68%
		Test	56,00%	56,00%	100,00%	71,79%
	32	Train	51,30%	64,00%	45,71%	53,33%
		Test	50,00%	61,54%	28,57%	39,02%
80:20	16	Train	63,64%	63,08%	100,00%	77,36%
		Test	45,45%	46,88%	93,75%	62,50%
	32	Train	40,15%	100,00%	3,66%	7,06%
		Test	51,51%	0,00%	0,00%	0,00%
90:10	16	Train	60,81%	60,81%	100,00%	75,63%
		Test	47,05%	47,05%	100,00%	63,99%
	32	Train	60,81%	60,81%	100,00%	75,63%
		Test	47,05%	47,05%	100,00%	63,99%

In Table 8, when using balanced data, the best-performing model in terms of testing was the one with an 80:20 ratio, achieving an accuracy of 51.85%, precision of 53.85%, recall of 50.00%, and an f1-score of 51.85. This best result is balanced with the accuracy of the training data, indicating no possibility of overfitting in the model. The results of each testing can be observed in the following table 8.

Table 8. Balance Data without Layer Bidirectional

Ratio	Batch Size	Data	Accuracy	Precision	Recall	F1-Score
70:30	16	Train	50,54%	0,00%	0,00%	0,00%
		Test	48,78%	56,00%	100,00%	71,79%
	32	Train	49,46%	49,46%	100,00%	66,19%
		Test	51,22%	51,22%	100,00%	67,74%
80:20	16	Train	59,81%	56,58%	81,13%	66,67%
		Test	51,85%	53,85%	50,00%	51,85%
	32	Train	50,47%	0,00%	0,00%	0,00%
		Test	48,15%	0,00%	0,00%	0,00%
90:10	16	Train	56,67%	54,12%	77,97%	63,89%
		Test	57,14%	62,50%	62,50%	62,50%
	32	Train	56,67%	54,12%	77,97%	63,89%
		Test	64,29%	66,67%	75,00%	70,59%

The information provided in Table 8 also reveals that the previous top-performing model, when using a 70:30 ratio, had a precision, recall, and f1-score of 0%. The confusion matrix of the model can be seen in Table 10.

Table 9. Confusion Matrix Result

	Actually Positive	Actually Negative
Predicted Positive	0%	51.22%
Predicted Negative	0%	48.78%

This indicates that the model was unable to correctly identify positive cases. These results imply that under sampling was not very successful in this scenario.

3.3.2. Test Analysis on Gated Recurrent with layer bidirectional

In this scenario, the model testing was performed using bidirectional layers and dividing the data into different ratios for each model: 90:10, 80:20, and 70:30. The data used in this case is imbalanced data, so it is important to consider the F1-score values for each model. Table 9 shows that when using imbalanced data with a 80:20 data split ratio and a batch size of 32, we obtained an accuracy of 57.58%, precision of 56.25%, recall of 56.25%, and an F1-score of 56.25%. The results of each test can be seen in the following table 9.

Table 10. Imbalance Data with Layer Bidirectional

Ratio	Batch Size	Data	Accuracy	Precision	Recall	F1-Score
70:30	16	Train	60,87%	60,87%	100,00%	75,68%
		Test	54,00%	55,10%	96,43%	70,13%
	32	Train	60,87%	60,87%	100,00%	75,68%
		Test	56,00%	56,00%	100,00%	71,79%
80:20	16	Train	65,15%	64,52%	97,56%	77,67%
		Test	54,55%	51,72%	93,75%	66,67%
	32	Train	62,88%	73,91%	62,20%	67,55%
		Test	57,58%	56,25%	56,25%	56,25%
90:10	16	Train	60,81%	60,81%	100,00%	75,63%
		Test	41,17%	42,85%	75,00%	54,54%
	32	Train	60,81%	60,81%	100,00%	75,63%
		Test	47,05%	47,05%	100,00%	63,99%

Based on Table 9, we can see that the influence of the bidirectional layer is very significant, as evidenced by the increase in percentages in all four values. The best model that utilizes the bidirectional layer consistently and effectively detects both positive and negative cases, compared to the best model without the bidirectional layer. It demonstrates a balanced and consistent performance. The result of the confusion matrix could be seen in Table 11.

Table 11. Best Model with Layer Bidirectional Confusion Matrix

	Actually Positive	**Actually Negative**
Predicted Positive	27.27%	21.21%
Predicted Negative	21.21%	30.30%

Based on the data in Table 7, out of 33 observed Twitter users, 9 users (27.27%) were detected as TP (True Positive), meaning they were classified as depressed and indeed experienced depression. Additionally, 10 users (30.30%) were detected as TN (True Negative), indicating that they were not diagnosed as depressed and did not experience depression. There were 7 users (21.21%) identified as FP (False Positive), meaning they were mistakenly classified as depressed when they actually did not experience depression. Furthermore, there were 7 users (21.21%) identified as FN (False Negative), indicating that they were not classified as depressed despite actually experiencing depression. The results of the predictions can be seen in table 12.

Table 12. Prediction Result

Username	**Tweets**	**Predict**
User1	'kangen', 'balas', 'balas', 'kakak', 'tertawa', 'pinya', 'kakak', 'teman', 'kak', 'tidur', 'asyik', 'lemah', 'hilang', 'kontak', 'ganti', 'kakak', 'teman', 'kak', 'tidur', 'kabar', 'kakak', 'ebae', 'invite', 'invite', 'eror', 'kali', 'angi', 'suka', 'musik', 'invite', 'kakak', 'confirm', 'daftar', 'lumayan', 'iseng', 'iseng', 'tambah', 'bagi', 'teman', 'balesanya', 'invite', 'kakak', 'confirm'	Negative
User2	'tikacemplung', 'bak', 'mandi', 'alhamdulillah', 'hidup', 'orang', 'suka', 'warga', 'tuban', 'uang', 'kaget', 'maaf', 'trauma', 'sejati', 'harap', 'manusia', 'widodo', 'anak', 'subuh', 'matkupil', 'tugas', 'susah', 'susah', 'uji', 'semester', 'terngaret', 'alam', 'semesta', 'ekle', 'revisi', 'bisa', 'twiteran', 'degdeganya', 'banget', 'tingkat', 'doir', 'tertawa', 'setel', 'lagu', 'cing', 'tarik', 'satarikna', 'tingkat', 'doir', 'tertawa', 'coba', 'motor', 'lembang', 'kayak', 'banget', 'jam', 'bagong', 'moga', 'gakyk', 'sekian', 'trimagaji', 'orang', 'dermawan', 'land', 'cruiser', 'jangkau', 'rover', 'kaliya', 'senak', 'takut', 'banget', 'jujur', 'jam', 'gin', 'sirine', 'ambulans', 'salah', 'kopi', 'dingin', 'hangat', 'abai', 'true', 'kep', 'harom', 'til', 'janah', 'sistur', 'ripiu', 'kak', 'dahla', 'gow', 'sajeu', 'minum', 'kopi', 'cantik', 'sulit', 'maaf', 'lahir', 'batin', 'tanggal', 'oktober', 'emosi', 'pilih', 'kasih', 'takut', 'banget', 'siklus', 'kemarin', 'jgnla', 'sakit', 'banget', 'sakit', 'manga', 'tampung', 'arianti'	Positive

Table 12 showed that user1 had previously been labeled as positive and user2 was also labeled as positive. However, after being predicted, the results revealed that user1 changed to negative, while user2 remained positive. This indicated that user2 truly experienced depression, while user1 did not actually experience depression.

4. DISCUSSION

In table 11, the results of the confusion matrix can be observed to be well-balanced, both in terms of true positive, true negative, false positive, and false negative values. It could be concluded that the model was effective in detecting both positive and negative cases. In other words, in contrast to the previous study conducted by Andre [10], who obtained true positive results of 58.82%, true negative of 11.76%, false positive of 5.88%, and false negative of 23.53%, the model from the current study showed effectiveness in detecting both types of cases. Similarly, in another research conducted by Putri [8], they obtained true positive results of 50.00%, true negative of 3.12%, false positive of 43.75%, and false negative of 3.12%. However, their model was effective in detecting only one type of case, either positive or negative. Therefore, it can be concluded that the model used in table 11 has an advantage as it effectively detected both types of cases in a well-balanced manner.

In the study, the model also obtained an accuracy of 54.54% on the testing data and an f1-score of 66.67%. Compared to Putri's research, the accuracy achieved was about 2% better when using GRU. However, it had a lower f1-score compared to using GRU, with a difference of about 5%. The results of Andre's research could not be directly compared to the GRU study, as the accuracy on the training data was 52.08%, and on the testing data, it was 70.59%. This indicated the presence of overfitting in the model used, making it unsuitable for a fair comparison with this study.

5. CONCLUSION

Based on the discussion in this research, several conclusions can be drawn. First, the GRU method with word2vec extraction can be used to identify depression among Twitter users. In this study, hyperparameter tuning was conducted with a train-test data ratio of 80:20, a batch size of 32, utilizing a bidirectional layer with 128 neurons. This resulted in an accuracy and f1-score of 62.88% and 65.15% on the training data, and an accuracy and f1-score of 54.54% and 66.67% on the test data, respectively. Second, the size of the dataset used can influence the model's performance. Under sampling the dataset led

to poor accuracy, precision, recall, and f1-score values. Third, the use of the bidirectional layer improved precision, recall, and f1-score values. Based on the previous discussion, the model was able to classify the dataset with a balanced distribution between positive and negative labels. For future research, it is recommended to expand the dataset and use a balanced dataset to achieve better model performance. Additionally, special attention should be given to the preprocessing of slang words in the Indonesian language, as it can introduce other disorders beyond depression, also known as non-depression.

REFERRENCES

[1] J. Sinuhaji, "Dunia Terisolasi Pandemi Covid-19, Pengguna Twitter Meningkat," *PikiranRakyat.com*. https://www.pikiran-rakyat.com/teknologi/pr-01634954/dunia-terisolasi-pandemi-covid-19-pengguna-Twitter-meningkat (accessed Dec. 02, 2022).

[2] M. A. Rizaty, "Pengguna Twitter di Indonesia Capai 18,45 Juta pada 2022," *DataIndonesia.id*. https://dataindonesia.id/digital/detail/pengguna-twitter-di-indonesia-capai-1845-juta-pada-2022 (accessed Dec. 06, 2022).

[3] P. Ananda, "WHO Sebut Pandemi Covid-19 Sebabkan Tingkat Depresi Naik 25%." https://www.okezone.com/tren/read/2022/03/15/620/2561866/who-sebut-pandemi-covid-19-sebabkan-tingkat-depresi-naik-25 (accessed Dec. 02, 2022).

[4] M. A. Rizaty, "Survei: 1 dari 3 Remaja Indonesia Punya Masalah Kesehatan Mental Artikel ini telah tayang di DSurvei: 1 dari 3 Remaja Indonesia Punya Masalah Kesehatan Mental," *dataindonesia.id*. https://dataindonesia.id/ragam/detail/survei-1-dari-3-remaja-indonesia-punya-masalah-kesehatan-mental (accessed Dec. 05, 2022).

[5] A. Pragholapati, "Covid-19 impact on students," 2020. doi: 10.35542/osf.io/895ed.

[6] K. Dianovinina and F. Psikologi, "Depresi pada Remaja: Gejala dan Permasalahannya Depression in Adolescent: Symptoms and the Problems," 2018.

[7] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," in *Proceedings - 2020 International Workshop on Electronic Communication and Artificial Intelligence, IWECAI 2020*, Institute of Electrical and Electronics Engineers Inc., Jun. 2020, pp. 98–101. doi: 10.1109/IWECAI50956.2020.00027.

[8] "Mengenali Depresi dari Cuitan Seseorang di Twitter," *nationalgeorgraphic.grid.id*. https://nationalgeographic.grid.id/read/13309312/mengenali-depresi-dari-cuitan-seseorang-di-twitter (accessed Dec. 05, 2022).

[9] J. Patihullah and E. Winarko, "Hate Speech Detection for Indonesia Tweets Using Word Embedding And Gated Recurrent Unit," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, p. 43, Jan. 2019, doi: 10.22146/ijccs.40125.

[10] G. Qorik, O. Pratamasunu, F. N. Fajri, D. Puji, and K. Sari, "Deteksi Tangan Otomatis Pada Video Percakapan Bahasa Isyarat Indonesia Menggunakan Metode Deep Gated Recurrent Unit (GRU)," 2022. [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[11] P. E. Sumolang and W. Maharani, "Depression detection on Twitter using Bidirectional long short term memory," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, pp. 369–376, 2022. doi:10.47065/bits.v4i2.1850.

[12] H. H. Windjatika and W. Maharani, "Depression detection on social media Twitter using long short-term memory," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 1835, 2022. doi:10.30865/mib.v6i4.4457.

[13] A. A. Simanungkalit, W. Maharani, and P. H. Gani, "Depression Detection on Twitter Social Media Platform using Bidirectional Long-Short Term Memory," *JINAV: Journal of Information and Visualization*, vol. 3, no. 2, pp. 190–203, Dec. 2022, doi: 10.35877/454ri.jinav1503.

[14] A. Nurdin, B. Anggo, S. Aji, A. Bustamin, and Z. Abidin, "PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS," *Jurnal TEKNOKOMPAK*, vol. 14, no. 2, p. 74, 2020.

[15] S. Kusumadewi and H. Wahyuningsih, "Model Sistem Pendukung Keputusan Kelompok untuk Penilaian Gangguan Depresii, Kecemasan dan Stress Berdasarkan DASS-42," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 2, p. 219, 2020, doi: 10.25126/jtiik.2020721052.

[16] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in English words," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 160–167. doi: 10.1016/j.procs.2019.08.153.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed

Representations of Words and Phrases and their Compositionality," Oct. 2013, [Online]. Available: http://arxiv.org/abs/1310.4546

[18] J. Chen, H. Jing, Y. Chang, and Q. Liu, "Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process," *Reliab Eng Syst Saf*, vol. 185, pp. 372–382, May 2019, doi: 10.1016/j.ress.2019.01.006.

[19] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[20] Q. Gu, L. Zhu, and Z. Cai, "Evaluation Measures of the Classification Performance of Imbalanced Data Sets," 2009.