OPEN ACCESS
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 10, No. 1, Jun 2024. pp. 62-77
doi: doi.org/10.21108/ijoict.v10i1.903

Sentiment Analysis on Presidential Election 2024 Using Word2Vec and Gated Recurrent Unit (GRU) with Genetic Algorithm Optimization

Syafa Fahreza [1], Erwin Budi Setiawan [2]

[1,2]*School of Computing, Telkom University*
Bandung, Indonesia

[*] syafafahreza@student.telkomuniversity.ac.id

Abstract

The evolution of social media has transformed platforms like Twitter from a mere information repository to a platform for expressing opinions and aspirations. Sentiment analysis on Twitter, particularly concerning the 2024 Indonesian presidential election, holds crucial importance for understanding public sentiment. The main contribution of this research is to optimize the Gated Recurrent Unit (GRU) model using Genetic Algorithm and combine feature expansion with Word2Vec for sentiment analysis on the topic of presidential election in Indonesia 2024. This research uses 39,791 datasets with GRU method, TF-IDF feature extraction, Word2Vec feature expansion with 142,545 corpus from IndoNews, and Genetic Algorithm optimization, the study achieves a peak accuracy of 86.46%, a 4.49% improvement over the baseline. By combining TF-IDF with a 5,000 maximum features, applying Word2Vec to the top 1 similarity, and utilizing Genetic Algorithm for feature optimization, this study demonstrates the effectiveness of these methods in improving the accuracy of sentiment analysis, thus significantly contributing to understanding public opinion during the 2024 Indonesian presidential election.

Keywords: Genetic Algorithm, GRU, Sentiment Analysis, TF-IDF, Word2Vec

I. INTRODUCTION

Advances in information technology have changed the role of social media from a mere repository of information to a platform for express opinions and aspirations [1]. One of the social media platforms that many people use is Twitter [2]. With 19.5 million users, Indonesia ranks fifth in the world for Twitter users, this information comes from Kominfo [3]. On Twitter, users can give opinions according to their conscience. Not all of these opinions are positive [1], so it is necessary to perform sentiment analysis to classify an opinion as positive or negative. Sentiment analysis on social media is becoming more important in the context of Presidential elections, which is one of the crucial phases in a country's democratic system [4].

Sentiment analysis is the process of identifying, understanding, and evaluating opinions, attitudes, or feelings contained in text, such as reviews, tweets, or social media posts. The main goal of sentiment analysis is to determine whether a text or opinion is positive or negative [5]. In this research, sentiment analysis will be conducted on the topic of the presidential election in Indonesia in 2024 on social media Twitter. Sentiments related to presidential candidates, policies, and relevant political issues reflect voters' views, hopes, concerns, and preferences. Through sentiment analysis on social media, we can understand people's perceptions and

Received on 13 Jan 2024. Revised on 25 Mar 2024. Accepted and Published on 30 Jun 2024.

responses to candidates and issues that arise during presidential election campaigns. This information provides valuable insights for the political, candidates and voters in understanding the evolving political dynamics and informing campaign strategies and policies that are more responsive to public needs.

Several deep learning techniques have been created for sentiment analysis, and one notable example is the Gated Recurrent Unit (GRU). One kind of artificial neural network design called GRU is used in Natural Language Processing (NLP) to simulate data sequences, such as text or human language data [7-9]. GRU has similarities with Long Short-Term Memory (LSTM), but with fewer parameters than LSTM [10]. In GRU, there is a gating mechanism that allows better information flow between recurrence cells, so that GRU can recall relevant information from the previous context in a more efficient way [11]. This makes GRU a good option in analyzing long texts with complex word sequences, such as social media posts.

This research uses Word2Vec as a feature expansion method and Genetic Algorithm as feature optimization. Word2Vec itself is used to convert data into vectors in the form of numbers, while Genetic Algorithm is used to solve optimization problems [12]. According to Mikolov, Word2Vec has several advantages, such Word2Vec can work quickly and efficiently for large data, is able to see and capture semantic relationships between words, and is able to generalize to other words [13]. On the other hand, the Genetic Algorithm has the advantage of effectively addressing extensive computational challenges. By combining GRU as a method, Word2Vec as feature expansion, and Genetic Algorithm as optimization expected can improve the accuracy of sentiment analysis model.

Previous research in sentiment analysis on Twitter is often limited in accuracy, data exploration, and model optimization. For example, [14] conducted sentiment analysis of Chinese e-commerce products using the GRU method, although obtaining an accuracy of 87%, the study has not utilized feature expansion techniques such as Word2Vec, resulting in limitations in word representation and lack of understanding of sentiment context. The lack of model optimization is also often an unresolved problem in previous studies. To address these issues, this research will combine GRU, TF-IDF, Word2Vec, and Genetic Algorithm. This research aims to improve model accuracy, enhance data representation through Word2Vec, and optimize model performance with Genetic Algorithm.

The main contribution of this research is to optimize the gated recurrent unit (GRU) model using genetic algorithm and combining feature expansion with word2vec for the 2024 presidential election in Indonesia. Knowing as much as the author does, no previous research has explored this combination, and the author sees the potential for increased accuracy through this integration. To produce a model with high accuracy, we used various methods, including GRU, TF-IDF, Word2Vec, and Genetic Algorithm. The research involved various scenarios, such as selecting the optimal ratio in the baseline, determining the maximum feature in feature extraction, selecting the corpus for feature expansion, and using optimization to increase the model's accuracy.

II. LITERATURE REVIEW

This research it is built upon a prior study. Research [15] carried out a survey of several deep learning techniques commonly applied to classification for sentiment analysis. In the study, the authors compared the LSTM, GRU, Bi-GRU, and Bi-LSTM algorithms using datasets taken from amazon reviews. The results obtained by Bi-GRU excel in accuracy with a value of 71.19%, slightly different from the accuracy obtained by GRU which is 71.06%. While GRU excels in precision, recall, and F1-score with values of 71%, 71.3%, and 71%, respectively.

Another study [16] was conducted with the aim of predicting gold prices with the aim of making it easier for people to obtain information about the future trends in gold market valuation. This study employs the Gated Recurrent Unit (GRU) technique for forecasting gold prices and adopts Mean Square Error (MSE) as a metric to gauge the accuracy of predictions. The outcomes indicated an MSE error rate of 0.111, an RMSE value of 0.334, and a coefficient of determination (R-squared) of 0.5.

In this [17] related work, sentiment analysis was performed on the SS-Tweet dataset using six classification methods that used both TF-IDF and N-Gram features. When compared to N-Grams, the research findings consistently showed that TF-IDF features performed better (3-4%). As a result, it can be said that TF-IDF is clearly a better option than N-Grams when it comes to using machine learning algorithms for text classification.

These results demonstrate how effectively TF-IDF can extract and represent important features needed for sentiment analysis from social media tweets and other datasets of similar type. Iqbal [18] conducted research using movie review dataset with Hindi language on sentiment analysis. The study compared 8 different model algorithms. Of all the algorithms tested the combination of GA-GRU got the best accuracy of 88.2%. The well-tuned hyperparameters keep the model from overfitting the dataset, which leads to an improvement in accuracy.

In the analysis presented by Research [14], sentiment analysis is conducted using the GRU approach. The dataset under consideration is derived from product reviews within the Chinese e-commerce sector. The overarching aim of this study is to assess and compare the performance of the GRU method with the LSTM method to see the potential for performance improvement. The accuracy results achieved by the GRU method on the Facebook dataset reached 87%. Meanwhile, research conducted by Farhan Wahyu Kurniawan and Warih Maharan [19] shows that differences in the architecture of the Word2Vec model have an impact on classification results. The outcomes demonstrated that applying a skip-gram model with a dimension of 100 resulted is the best classification performance. The precision of the model reached 64.4%, with a recall of 58%, and an F1-score of 61.1%.

III. RESEARCH METHOD

A. Design System

In general, the system aims to construct a sentiment analysis model designed to categorize and analyze user opinions. Sentiment analysis model developed in this research consists of several steps including Data Crawling, Labeling Data, Data Preprocessing, Extraction Feature, Expansion Feature, Classification, Sentiment Prediction, and Evaluation. Figure 1 provides a system overview made in this study.

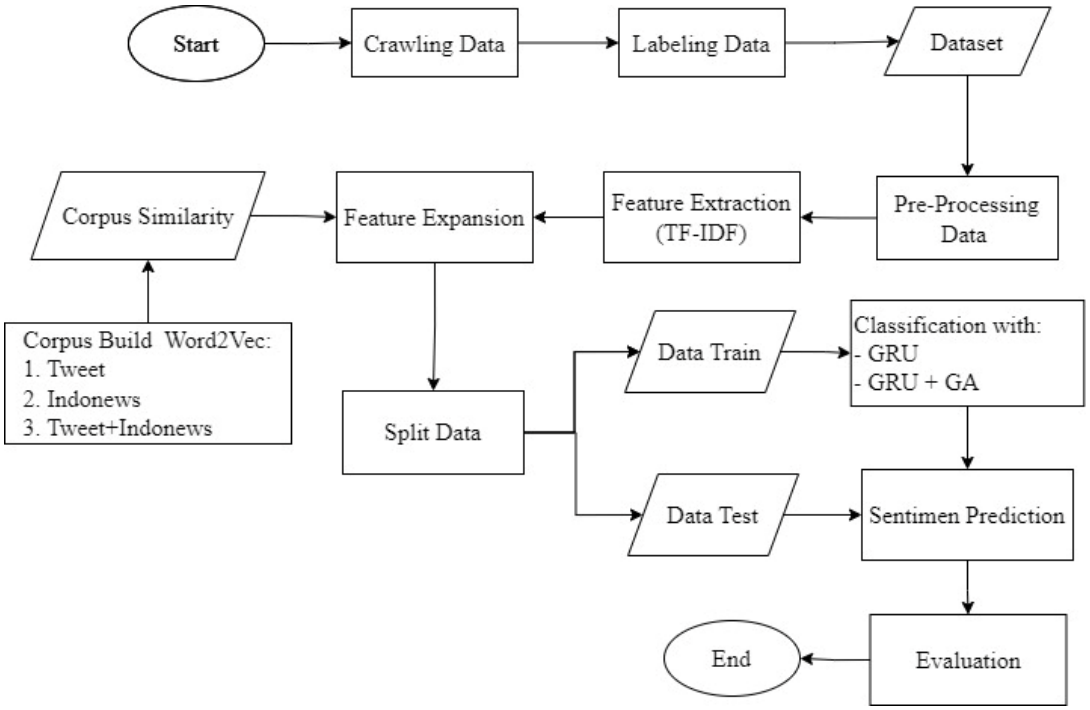

Fig. 1. System Overview

B. Crawling Data

Finding data from a certain source is known as "crawling.", the data used in this study came from Twitter social media platform. The data taken is data with keywords Calon Presiden, Capres, Anies Baswedan, Prabowo Subianto, and Ganjar Pranowo. 37,391 data were obtained from the 5 keywords. The results of crawling data will be automatically stored in during the crawling process using the python programming language.

TABLE I
KEYWORD DATA CRAWLING

Keyword	Amount	Ratio (%)
Anies Baswedan	10,434	27.90
Ganjar Pranowo	8,027	21.47
Capres	7,296	19.51
Calon Presiden	6,972	18.65
Prabowo Subianto	4,662	12.47
Total	37,391	100

C. Labeling Data

In the process of developing datasets for classification systems, steps are needed to ensure that the collected datasets have accurate class labels. Therefore, the results of crawling data will be labeled. This step is important to ensure that each data or sample in the dataset has a class identification that matches its characteristics, thus allowing the classification system to learn and generalize patterns better.The labels will be given in 2 classes which are, positive and negative. The data will be labeled manually by 3 people using the concept of majority votes. This means that decisions are made based on the choice that gets the most number of votes compared to other choices.

1) The positive class (1) indicates that the tweet has a positive sentiment such as happiness and approval.
2) The negative class (-1) indicates that the tweet has a negative sentiment such as anger and rejection.

TABLE II
LABELED DISTRIBUTION

Label	Amount	Ratio (%)
Positive	21,866	58.48
Negative	15,525	41.52
Total	37,391	100

TABLE III
LABELING EXAMPLE

Sentiment	Label
@Melihat_Indo @ganjarpranowo Ayo generasi muda bangsa ini, kita tunjukkan kekompakan kita mendukung calon presiden yang hebat yakni Pak Ganjar Pranowo	1
@devcepi60 @BradHarizz @ganjarpranowo Emang orang yang tak layak dipilih kok. Sekali lagi: tidak layak. Selamatkan bangsa dengan tidak memilih capres yang tidak layak.	-1

D. Preprocessing Data

Data preprocessing is a process where raw data is cleaned so that the data is ready to be used for further analysis. At this stage, the data is ensured clean to avoid prediction errors that will be made. Data retrieved from Twitter frequently includes non-relevant characters, such as symbols, emoticons, mentions, and links, which may not contribute significantly to the classification process. Therefore, a process is needed to clean and fix the problem.

The steps in preprocessing are as follows:
1) ***Data Cleaning****:* Data cleaning is the process of removing noise in data. In real life, data frequently contains noise and inconsistencies. In the context of tweet data cleaning, steps are taken to remove URLs, numbers, symbols, and attributes that contain missing values or blanks.

TABLE IV
DATA CLEANING

Before	After
@Melihat_Indo @ganjarpranowo Ayo generasi muda bangsa ini, kita tunjukkan kekompakan kita mendukung calon presiden yang hebat yakni Pak Ganjar Pranowo	Ayo generasi muda bangsa ini kita tunjuk kompak kita mendukung calon presiden yang hebat yakni Pak Ganjar Pranowo

2) ***Case Folding****:* Case folding is the process of converting text in dataset into lowercase consistently. The purpose of case folding is to eliminate the difference in case (uppercase and lowercase) in the text.

TABLE V
CASE FOLDING

Before	After
Ayo generasi muda bangsa ini kita tunjuk kompak kita dukung calon presiden yang hebat yakni Pak Ganjar Pranowo	ayo generasi muda bangsa ini kita tunjuk kompak kita mendukung calon presiden yang hebat yakni pak ganjar pranowo

3) ***Stopword Removal****:* The process of stopword removal entails eliminating common words that lack specific meaning and do not contribute significant information to a text or document. Stopword removal helps simplify text representation and improve accuracy in text analysis or language modeling.

TABLE VI
STOPWORD REMOVAL

Before	After
ayo generasi muda bangsa ini kita tunjuk kompak kita dukung calon presiden yang hebat yakni pak ganjar pranowo	generasi muda bangsa kompak mendukung calon presiden hebat ganjar Pranowo

4) ***Stemming****:* Stemming involves the elimination of word affixes in order to derive the base form of a word. The stemming process can help reduce word variation and simplify matching in text analysis.

TABLE VII
STEMMING

Before	After
generasi muda bangsa kompak mendukung calon presiden hebat ganjar Pranowo	generasi muda bangsa kompak dukung calon presiden hebat ganjar Pranowo

5) ***Tokenization****:* Tokenization is the process of splitting text or sentences into smaller units called tokens. Tokens can be words, phrases, symbols, or any other unit that is relevant in the context of natural language processing.

TABLE VIII
TOKENIZATION

Sebelum	Sesudah
generasi muda bangsa kompak dukung calon presiden hebat ganjar Pranowo	['generasi', 'muda', 'bangsa', 'kompak', 'dukung', 'calon', 'presiden', 'hebat', 'ganjar', 'Pranowo']

E. Feature Extraction

Feature extraction plays an important role in document processing in search engines as it has a significant impact on the success of the text mining process. If the results of feature extraction are not accurate, the information found in text mining will not match the expected criteria. As a result, the search results displayed by the search engine will not be in accordance with the user's wishes [20]. In this research, TF-IDF (Term Frequency-Inverse Document Frequency) is used as a feature extraction method. TF-IDF was chosen as feature extraction because it has several advantages. First of all, TF-IDF is able to retrieve the frequency of words in a document, which makes it possible to evaluate the importance of those words in the context of a particular document. Words that appear more frequently in a document are considered more important and are given a higher weight, while words that appear less frequently are given a lower weight. In addition, TF-IDF is also able to help reduce the impact of common words that appear in many documents, which may not provide useful information in identifying or distinguishing documents [21]. Feature extraction using TF-IDF can be calculated using the following equation:

$$tf_t = \frac{D}{df_t} \tag{1}$$

$$idf_t = \log\left(\frac{D}{df_t}\right) \tag{2}$$

$$W_{dt} = tf_{dt} \times idf_{dt} \tag{3}$$

In the TF-IDF calculation process, the first step is to determine the Term Frequency (TF) quantifies the frequency of a term within a document, determined by dividing the number of times the term appears by the total number of words in the document. For example, if the first document has 1000 words and the term "intelligence" appears 5 times, then the TF for the first document is 0.005. The second step is to calculate the Inverse Document Frequency (IDF), which measures how common or rare a term is across the document set. IDF is computed by dividing the total number of documents by the number of documents that contain the term, and then applying the natural logarithm to the result. In this example, since the term "intelligence" appears in all five documents, its IDF value is 0. This indicates that the term provides no useful information in distinguishing the documents. The last step is to multiply the TF value by the IDF value to get the TF-IDF value. In this case, since the IDF for the term "intelligence" is 0, the TF-IDF value for the term in all documents will be 0. This indicates that the term does not provide significant weight or relevance in distinguishing the documents in the collection.

F. Feature Expansion

Word2Vec is one of the word embedding methods improved by Mikolov [22]. The purpose of Word2Vec is to represent words into a vector of length N, so that it can be used to identify the relationship or correlation between words.Word2Vec has 2 neural network architectures namely skip-gram and Continuous Bag of Words (CBOW). The difference between the two is that skip-gram architecture aims to predict words around the word being processed (input), while CBOW aims to predict words (output) when given context or words around it (input). Based on research conducted by Mikolov, the results obtained, when using large data CBOW will perform better than skip-gram, while skip-gram will perform better when used to learn new words [23].

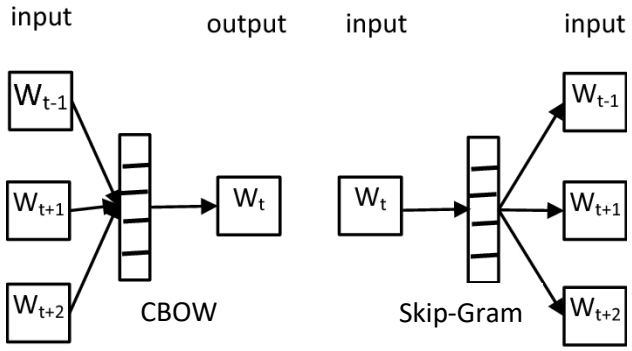

Fig. 2. Word2Vec Architecture

So it can be concluded that both of them have their respective advantages, CBOW will work better when applied to news articles while skip-grams have better better performance when applied to tweets. Because the data used in this research is twitter data, the type of Word2Vec used is skip-grams. First, Word2Vec is trained using tweet data and Indonews to create a similarity corpus, akin to the Word2Vec approach, enabling the model to capture word similarities. Training on this corpus allows Word2Vec to build vector representations of words that reflect the context and semantic relationships within the Indonesian language. This approach aims to enhance the understanding of words in their respective contexts, enabling the model to measure word similarity based on similar patterns of usage.

In this research, the Word2Vec model will be trained on an Indonews corpus to generate a word vector representation, which will be the input for the GRU (Gated Recurrent Unit) model. Using the word vector representation, the GRU model can learn complex patterns in the text, such as relationships between words and contextual meanings, thus improving its capabilities in natural language processing tasks such as text classification, translation, and text generation. The corpus will be built using the Word2Vec Skip-gram model for word embedding. This corpus contains words organized by their degree of similarity, from highest to lowest. In this research, the highest accuracy was obtained when using the corpus from Indonews. The following is an example of the top 10 similarity of the corpus that has been achieved using word2vec.

TABLE IX
WORD2VEC SIMILARITY EXAMPLE

Teks	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
Presiden	jokowi	widodo	negara	pemerintah	calon
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	dukung	pemimpin	kepala	parlemen	eksekutif

For example, there is a tweet "Presiden negara mengumumkan kebijakan baru tentang pembangunan infrastruktur." Then, with feature expansion using the top similarity 10 feature size using the Indonews corpus, in the TF-IDF vector representation, the word "negara" has zero weight. Since "negara" is in the top 10 similarity of the word "presiden", the word "negara" will have the same weight as the weight of "presiden". For the top 10 similarity of the word "presiden" in the Indonews corpus, can be seen in table IX.

G. Gated Recurent Unit (GRU)

GRU (Gated Recurrent Unit) represents a specific type of recurrent neural network (RNN) architecture utilized in the fields of machine learning and natural language processing. GRU was developed as a lighter and more efficient than LSTM (Long Short-Term Memory) [11]. In theory, GRU can be trained faster than LSTM because it involves fewer calculations [23].

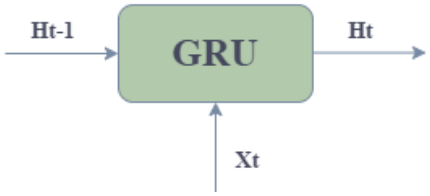

Fig. 3 GRU Architecture

In contrast to LSTM, which incorporates four gates, GRU is characterized by having only two gates, specifically the reset gate and the update gate. The update gate plays a role in helping the model to determine the extent to which past information (from the previous timestamp) should be forwarded to the future. While the reset gate is used by the model to determine how much past information should be ignored.

Its functioning involves utilizing the input (Xt) and the hidden state (Ht-1) from the preceding timestamp at each time step (t). This input is then processed to generate a new hidden state output (Ht), which is subsequently passed on to the next timestamp in the sequence.

H. Genetic Algorithm

Genetic Algorithm is one of the method used to resolve issues related to optimization. Genetic Algorithm was first developed by John Holland by adopting the basic theory described by Charles Darwin. The theory explains that in the process of natural evolution, each individual must be able to adapt to the surrounding environment in order to survive [24]. In the context of Genetic Algorithm, the population of individuals is represented by chromosomes that represent potential solutions to a problem [24]. The evolutionary process in Genetic Algorithms involves various operations such as selection, reproduction, crossover, and mutation. Through repeated iterations, the best individuals that are able to adapt to the environment and have an advantage in terms of fitness will have a greater chance of being passed on to the next generation [25].

In applying genetic algorithm (GA) to optimize GRU hyperparameters, a number of predefined parameters are required. These parameters include the range of possible values for each hyperparameter such as the number of neurons, learning rate, and dropout rate. In addition, it is also necessary to determine the performance evaluation function that will be used to assess each individual in the population. The GA process begins with the creation of an initial population consisting of a number of individuals representing a random combination of hyperparameters. Next, the individuals are evaluated using a performance evaluation function that is usually based on the performance of the GRU model on a validation dataset or through cross-validation. After that, individuals in the population are selected to perform genetic operations such as selection, crossover, and mutation. Selection is based on individual performance, where the best performing individuals have a greater chance of being retained in the next population. Crossover is used to create new genes by combining good traits from selected parents.

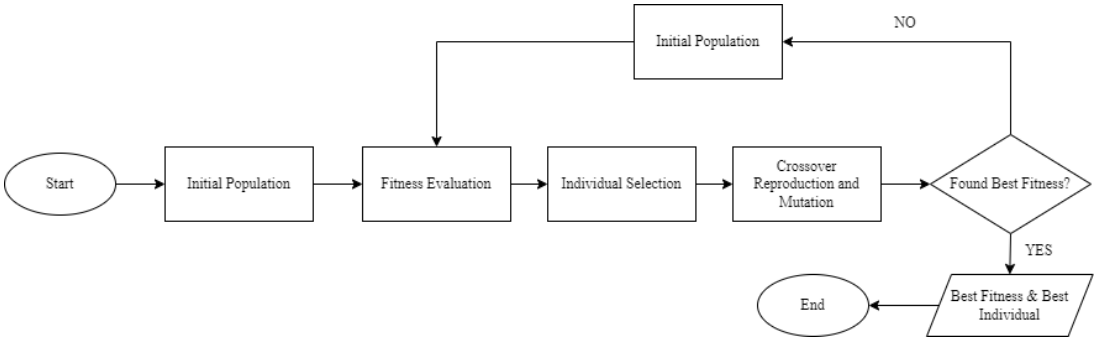

Fig. 4 Genetic Algorithm Flowchart

Mutations are performed to introduce new variations into the population and prevent premature convergence to a sub-optimal solution. This process is repeated in multiple generations until certain stopping criteria are met, such as reaching the maximum number of generations or no performance improvement in several consecutive generations. The final result of the genetic algorithm is the combination of hyperparameters that gives the best performance for the GRU model on the specified task.

I. GRU+GA

The concept of combining a genetic algorithm (GA) with the GRU model occurs in the process of finding the optimal parameters for the GRU architecture. In this context, GA acts as an optimization mechanism that searches for the best combination of parameters for the GRU architecture. Specifically, the GA adjusts the parameters that affect the GRU architecture, including the number of GRU units, dropout rate, and other relevant parameters. In each generation, the GA generates individuals that represent various combinations of the values of these parameters. These individuals are then evaluated using a fitness function, which in this context measures the performance of the GRU model built using the parameters it represents. Through the process of selection, crossover, and mutation, these individuals then interact to create the next generation of to create the next generation of parameters which may be better in terms of the performance of the GRU model. Thus, GA effectively guides the search for optimal parameters for the GRU architecture, which can result in a better GRU model in a given classification task. Thus, the merging of GA and GRU occurs in the process of searching and optimizing the the parameters of the GRU architecture to improve the overall performance of the model. overall model performance.

In combining GRU (Gated Recurrent Unit) and Genetic Algorithm (GA), there are several fundamental reasons that strengthen these two approaches. First of all, GRU was chosen as the recurrent model due to its ability to handle the vanishing gradient problem, which is often experienced by LSTM models. The advantage of GRU, which has a simple structure with two gates (update and reset), makes it efficient and effective in long-term learning. The selection of GA as an optimization algorithm is in line with the objective to find the optimal parameters for the GRU architecture. GA enables parameter optimization by performing exploration and exploitation in the parameter search space. The evolutionary process implemented by GA can adjust the number of GRU units, dropout rate, and other parameters, ensuring the GRU model can adapt itself to a given classification or prediction task.

J. Evaluation

Performance measures are carried out to evaluate the model model that will be built. The evaluation is performed using a confusion matrix to compare model predictions with the actual value of the data used for evaluation.

TABLE X
CONFUSION MATRIX

Confusion Matrix		Actual Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

In the table there are four terms, each of which has the following meaning as follows: TN (True Negative) is the number of correct predictions that a negative data is predicted to be negative, FN (False Negative) is the number of incorrect predictions that data is predicted to be negative, FP (False Positive) is the number of wrong

predictions that that a data is predicted to be positive, and TP (True Positive) is the number of correct predictions that a data is predicted to be positive [26]. Here are some of the calculations used in the confusion matrix calculation:

1) Accuracy: Accuracy gauges the degree to which a model correctly classifies data. Accuracy is expressed as a percentage of the number of correct predictions compared to the total amount of data [26]. Accuracy can be calculated with the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

2) Precision: Precision measures the extent to which the positive predictions given by the model are correct. Precision is defined as the percentage of correctly predicted positive cases in relation to the total number of predicted positive cases [26]. Precision can be calculated by the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

3) Recall: Recall, which is often referred to as Sensitivity or True Positive Rate, quantifies how effectively the model can identify positive situations in general. Recall is calculated as the percentage of correctly predicted positive cases in relation to the total number of true positive cases [26]. Recall can be calculated with the following formula:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

4) F1-Score: The F1-score serves as a comprehensive metric by combining precision and recall, thereby offering a balanced evaluation of a model's performance. Precision is comparison of positive correct predictions compared to the overall positive predicted positives [26]. F1-score can be calculated with the following formula:

$$F1 - Score = \frac{2\ x\ (Precision\ \times Recall)}{(Precision + Recall)} \tag{7}$$

IV. RESULTS AND DISCUSSION

In this study, several steps of testing scenarios were applied to find the optimal performance results. These steps were executed with the expectation that accuracy and F1-score would increase with each phase of testing performed. This research has four test cases. In the first scenario, the ratio of train data to test data is compared in order to establish the baseline for the GRU model. Using TF-IDF feature extraction on the baseline is the second scenario. In the third scenario, Word2Vec is used to expand features. For the fourth scenario is using a Genetic Algorithm for optimization. The average of the five test outcomes determines the accuracy rating and F1 score for each scenario.

A. Scenario 1 Testing Best Split Size

In the first scenario, we compare the accuracy and F1-score of three data splits: 70:30, 80:20, and 90:10. The main purpose of this scenario is to find the right data sharing ratio between the training set and the testing set. This ratio can significantly affect the performance of the model. Moreover, this test can also help prevent overfitting and underfitting [27]. This process consists of training the model on the training and measuring the performance on the testing data. After evaluation, the data split size that gives the best performance will be

selected. The best result will be implemented in the next scenario, which is TF-IDF (Term Frequency-Inverse Document Frequency). Table 5 displays the outcome achieved from scenario 1.

TABLE XI
RESULTS OF SCENARIO 1

Split Size	Accuracy (%)	F1-Score (%)
90:10	**81.97**	**81.47**
80:20	81.50	81.00
70:30	80.66	79.94

The test results indicate that the optimal split for training and test data is achieved with a division of 90% for training and 10% for testing. This particular data split provides the highest accuracy and F1-score, reaching 81.97% for accuracy and 81.47% for F1-score. These values will be used as the baseline for the upcoming scenario.

B. Scenario 2 Testing Best Max Feature

In the second stage scenario, following the attainment of the initial baseline accuracy through GRU classification, attention is directed to identifying the most effective maximum features through TF-IDF feature extraction. The purpose of this scenario is to determine the relationship between max features and model performance based on accuracy and F1-score. The maximum features to be compared are 1,000, 2,000, 5,000, 8,000, and 12,000. The following table will show the comparison between the maximum features that have been tested.

TABLE XII
RESULTS OF SCENARIO 2

Max Feature	Accuracy (%)	F1-Score (%)
1000	81.71 (-0.26)	81.31 (-0.16)
2000	82.66 (+0.69)	82.36 (+0,89)
5000	**82.96** **(+0.99)**	**82.57** **(+1,1)**
8000	82.95 (+0.98)	82.56 (+1.09)
12000	82.90 (+0.93)	82.41 (0.94)

From the data in the table, it can be inferred that the maximum feature of tf-idf feature extraction is 5000. From the max feature, the accuracy of 82.96% is obtained, increasing 0.99% from the baseline accuracy and for the F1-Score is 82.57%, increasing 1.1% from the baseline. Furthermore, these results will be used for the next scenario using the Word2Vec feature expansion.

C. Scenario 3 Testing Top Similarity

Testing scenario 3 enhances the best outcomes from scenario 2 by applying the Word2Vec feature expansion. Three different corpuses—Twitter, IndoNews, and Twitter + IndoNews—were employed to execute the Word2Vec feature expansion. The corpus comprises a set of words sharing comparable meanings. In Scenario 3, the top 1, top 5, and top 10 similarity scores for every word across all corpuses are compared.

TABLE XIII

RESULTS OF SCENARIO 3

Top-N	Twitter		IndoNews		Twitter+IndoNews	
	Accuracy (%)	**F1-Score (%)**	**Accuracy (%)**	**F1-Score (%)**	**Accuracy (%)**	**F1-Score (%)**
1	82.13 (+0.16)	81.68 (+0.21)	**83.00 (+1.03)**	**82.70 (+1.23)**	82.76 (+0.79)	82.38 (+0.91)
5	81.50 (-0.47)	81.03 (-0.44)	82.70 (+0.73)	82.26 (+0.79)	82.47 (+0.5)	82.06 (+0.59)
10	79.95 (-2.02)	79.57 (-1.9)	82.14 (+0.17)	81.67 (+0.2)	81.94 (-0.03)	81.55 (+0.08)

By using the three corpus, the accuracy and F1 score increase from the baseline. However, of the three corpuses that have been created, the best accuracy and F1-Score are obtained when using the corpus from IndoNews. With an increase in accuracy of 1.03% and F1 Score 1.23% from baseline at top 1. This highlights the importance of choosing the right corpus, as the IndoNews dataset seems to capture valuable language patterns that greatly contribute to better predictions. The improvement suggests that the IndoNews corpus not only aligns well with the model's goals but also captures subtle language nuances, resulting in more accurate and reliable predictions.

D. Scenario 4 Using Genetic Algorithm

In this fourth scenario is the use of Genetic Algorithm as feature optimization. In each generation of a Genetic Algorithm (GA), various tasks such as crossover, mutation, generation of offspring, and fitness evaluation are performed. The time complexity of the GA is influenced by the characteristics of the fitness function used. The table below shows the parameters in the Genetic Algorithm used in the test.

TABLE XIV
EVOLUTIONARY GA PARAMETER

Evolutionary Parameter	Value
Cxpb	0.5
Mutpb	0.2
Ngen	5

TABLE XV
GENETIC GA PARAMETER

Genetic Parameter	Value
Mate	Indpb = 0.5
Mutate	Mu = 0, Sigma = 1, indpb = 0.2
Select	Tournsize = 3

In the context of the evolutionary GA parameters applied, a Cxpb (crossover probability) value of 0.5 indicates that each pair of individuals has a 50% chance of crossover in each generation. Meanwhile, the Mutpb (mutation probability) value of 0.2 indicates that each gene in an individual has a 20% probability of mutating. With Ngen (number of generations) of 5, the algorithm will run for five generations before stopping, with crosses and mutations occurring according to the probabilities set for each operation.

For the applied genetic algorithm parameters, in the mate process, the Indpb value of 0.5 indicates the probability of each pair of individuals to experience crossover. In the mutate process, the values Mu = 0 and Sigma = 1 indicate the mean and standard deviation of the normal distribution used to generate mutation values, with a mutation probability indpb of 0.2 for each gene in the individual. In the selection process (select), Tournsize = 3 indicates the tournament size used in selecting individuals for reproduction in each generation.

From the test results, the output shows the best individual (chromosome) and the best fitness value. In this case the best individuals are considered as best units and the best fitness value is considered as accuracy. The best unit obtained is 78 with a drop out value 0.21 and the accuracy obtained is 86.46% increased 4.49% from baseline.

E. Discussion

The test results in this study provide a deep understanding of the effectiveness of the GRU model in performing sentiment analysis on Twitter data with the topic of the presidential election in Indonesia in 2024. In this study, we will compare several scenarios to see whether or not each scenario has an effect on the increase in accuracy of the model. The first scenario shows that the selection of different split data can affect the results. The accuracy and F1 value of the results from the best split data from this scenario are considered as a baseline for further comparison. The variation in accuracy across different data splits shows that proper selection of data splits is crucial to obtain consistent and accurate results. This can be explained by the uneven distribution of data, where some split data may contain more instances of a particular sentiment than others.

The second scenario explores the use of TF-IDF as a feature extraction method. Tests on the maximum feature variation have been conducted with a range from 1,000 to 12,000 features. In the test with a maximum of 5,000 features, the accuracy and F1-score reached the highest values, increasing by 0.99% for accuracy and 1.1% for F1-score. However, when testing at a maximum of 12,000 features, the accuracy decreased with an increase ratio of only 0.93% for accuracy and 0.94% for F1-Score. This improvement in accuracy with TF-IDF shows that proper feature selection can improve model performance. TF-IDF helps the model to focus on the most relevant and informative words in the data.

In the third scenario, the application of Word2Vec as feature expansion illustrates that corpus selection has an impact on accuracy results. The use of the IndoNews corpus on similarity top 1 resulted in the highest accuracy, with an increase of 1.03%, along with an increase in F1-Score of 1.23%. This shows that the word vector representation can help the model to understand the contextual meaning of words. The use of the IndoNews corpus shows that a word vector representation specific to Indonesian can improve model performance. The last scenario involves the use of Genetic Algorithm as a feature optimization approach. The results show that the implementation of Genetic Algorithm can improve the accuracy of the model by producing the best unit, best dropout, and best fitness. This shows that feature optimization using Genetic Algorithm can help the model to find the best combination of features to improve performance.

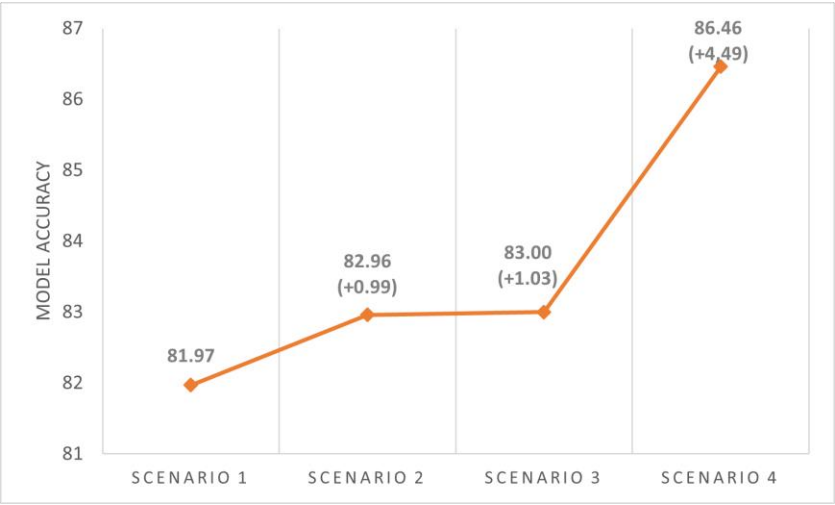

Fig. 5 Testing Result for All Scenario

This research shows that combining the GRU model with TF-IDF, Word2Vec, and Genetic Algorithm is effective for analyzing public sentiment towards the presidential election in Indonesia. These results can be a reference for further research in sentiment analysis of Twitter data. In addition, this model can be used by political organizations to monitor public opinion towards the presidential election, survey agencies to improve the accuracy of their surveys, and media companies to understand public sentiment towards their products and services. This model has the potential to provide valuable insights for various parties who want to understand public opinion in Indonesia. However, this study needs to be followed up with some further research. First, the performance of the GRU model needs to be tested on Twitter data with other topics. Second, the effect of GRU model parameters on the analysis results needs to be studied further. Third, the representation of word vectors from other corpus needs to be analyzed to improve the accuracy of the model. Lastly, the performance of Genetic Algorithm needs to be compared with other feature optimization methods. Further research is important to strengthen the findings of this study and improve the effectiveness of the GRU model in analyzing the sentiment of Twitter data.

V. CONCLUSION

In this research, an analysis of sentiments was carried out using a dataset from Twitter social media with a focus on the topic of the president. The dataset used consists of 37,391 to be labeled as negative and positive, which is performed through a manual labeling process. Before being used as a dataset for the model, a preprocessing stage was carried out to make it easier for the model to understand and process the data. This research involves testing using four scenarios on the GRU model. The first scenario compares the test results using different split data, where the best result is used as a baseline for comparing with the next test. The second scenario involves the use of TF-IDF as the feature extraction method, with testing on a variety of max features covering the range of 1000 to 12000. The best results were achieved when the max feature set was 5,000. The third scenario involves the use of Word2Vec as feature expansion, with testing on three different corpus: Twitter, IndoNews, and Twitter+IndoNews corpus combined. Each corpus was tested for top 1, top 5, and top 10 similarity values. The last scenario involves applying a Genetic Algorithm as feature optimization, which determines the best unit, best dropout, and best fitness. The test results show that the highest accuracy achieved is 86.46%, which shows an improvement of 4.49% compared to the baseline. This performance was achieved when combining TF-IDF with 5,000 maximum features, applying Word2Vec with 142,545 corpus from IndoNews at top 1 similarity, and applying Genetic Algorithm for feature optimization.

REFERENCES

[1] N. A. Shafirra and I. Irhamah, "*Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi* Speech-to-Text (STT) *Menggunakan Metode* Convolutional Neural Network (CNN)," J. Sains dan Seni ITS, vol. 9, no. 1, 2020, doi: 10.12962/j23373520.v9i1.51825.

[2] Kamil, G., & Setiawan, E. B. (2023). Aspect-Level Sentiment Analysis on Social Media Using Gated Recurrent Unit (GRU). *Building of Informatics, Technology and Science (BITS)*, *4*(4), 1837-1844.

[3] www.kominfo.go.id, "Indonesia *Peringkat 5 Pengguna Twitter*" 2022. https://www.kominfo.go.id/content/detail/2366/%20indonesia-peringkatlima-penggunatwitter/0/sorotan_media (accessed Apr. 20, 2023).

[4] Kurniawan, I., & Susanto, A. (2019). *Implementasi Metode K-Means dan NaÃ ve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019*. Jurnal Eksplora Informatika, 9(1), 1-10.

[5] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107134.

[6] Kanojia, D., & Joshi, A. (2023). Applications and Challenges of Sentiment Analysis in Real-life Scenarios. *arXiv preprint arXiv:2301.09912*.

[7] Huang, Z., Yang, F., Xu, F., Song, X., & Tsui, K. L. (2019). Convolutional gated recurrent unit–recurrent neural network for state-of-charge estimation of lithium-ion batteries. *Ieee Access*, *7*, 93139-93149.

[8] Santur, Y. (2019, September). Sentiment analysis based on gated recurrent unit. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-5). IEEE.

[9] Al Wazrah, A., & Alhumoud, S. (2021). Sentiment analysis using stacked gated recurrent unit for arabic tweets. *IEEE Access*, *9*, 137176-137187.

[10] Hidayatullah, A. F., Cahyaningtyas, S., & Hakim, A. M. (2021, February). Sentiment analysis on twitter using neural network: Indonesian presidential election 2019 dataset. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1077, No. 1, p. 012001). IOP Publishing.

[11] Xing, Y., & Xiao, C. (2019, August). A GRU Model for Aspect Level Sentiment Analysis. In *Journal of Physics: Conference Series* (Vol. 1302, No. 3, p. 032042). IOP Publishing.

[12] F. W. Kurniawan and W. Maharani, "Indonesian Twitter Sentiment Analysis Using *Word2Vec*," 2020 Int. Conf. Data Sci. Its Appl. ICoDSA 2020, pp. 31–36, 2020, doi: 10.1109/ICoDSA50139.2020.9212906.

[13] Rezeki, S. R. I. (2020). *Penggunaan Sosial Media Twitter dalam Komunikasi Organisasi (Studi Kasus Pemerintah Provinsi Dki Jakarta Dalam Penanganan* Covid-19). Journal of Islamic and Law Studies, *4*(2).

[14] J. S. Lee, D. Zuba, and Y. Pang, "Sentiment analysis of Chinese product reviews using gated recurrent unit," Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol., pp. 173–181, 2019, doi: 10.1109/BigDataService.2019.00030.

[15] Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment analysis using gated recurrent neural networks. *SN Computer Science*, *1*, 1 13.

[16] Alkahfi, I., & Chiuloto, K. (2021). *Penerapan Model Gated Recurrent Unit Pada Masa Pandemi Covid-19 Dalam Melakukan Prediksi Harga Emas Dengan Menggunakan Model Pengukuran* Mean Square Error. *Snastikom Ke*, *8*, 225-232.

[17] Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341-348

[18] Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access*, *7*, 14637-14652.

[19] F. W. Kurniawan and W. Maharani, "Indonesian Twitter Sentiment Analysis Using *Word2Vec*," 2020 Int. Conf. Data Sci. Its Appl. ICoDSA 2020, pp. 31–36, 2020, doi: 10.1109/ICoDSA50139.2020.9212906

[20] Bania, R. K. (2020). COVID-19 public tweets sentiment analysis using TF-IDF and inductive learning models. *INFOCOMP Journal of Computer Science*, *19*(2), 23-41.

[21] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," in Procedia Computer Science, 2019, vol. 152, pp. 341–348. doi: 10.1016/j.procs.2019.05.008.

[22] Savytska, L. V., Vnukova, N. M., Bezugla, I. V., Pyvovarov, V., & Sübay, M. T. (2021). Using Word2vec technique to determine semantic and morphologic similarity in embedded words of the Ukrainian language.

[23] Zaman, L., Sumpeno, S., & Hariadi, M. (2019). *Analisis Kinerja LSTM dan GRU sebagai Model Generatif untuk Tari Remo*. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, *8*(2), 142-150.

[24] Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia tools and applications*, *80*, 8091-8126.

[25] Pratiwi, B. P., Handayani, A. S., & Sarjana, S. (2020). *Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix*. *Jurnal Informatika Upgris*, *6*(2).

[26] A. Suresh, "What is a confusion matrix?," 2020. https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (accessed May 14, 2023).

[27] Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules, 26(4), 1111.