Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 511-519

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.855
p-ISSN: 2723-3863
e-ISSN: 2723-3871

FINE-GRAINED SENTIMENT ANALYSIS IN SOCIAL MEDIA USING GATED RECURRENT UNIT WITH SUPPORT VECTOR MACHINE

Wida Sofiya[1], Erwin Budi Setiawan[*2]

[1,2]Informatics, Faculty of Informatics, Universitas Telkom, Indonesia
Email: [1]wydaun@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

(Article received: January 24, 2023; Revision: February 14, 2023; published: June 26, 2023)

Abstract

Social media platforms are widely used to share opinions, leading to a large growth of text data on the internet. This data can be a key source of up-to-date and inclusive information by conducting sentiment analysis. Typically, sentiment analysis research classifies binary based on the polar values generated. However, this has its limitations, such as classifying sentences containing positive and negative expressions, leading to incorrect predictions. Fine-grained sentiment analysis provides more precise results by associating values with more than two classification targets. The objective of this study is to carry out sentiment analysis at a fine-grained level related to public policy in Indonesia using the GRU-SVM model with feature extraction and expansion techniques. However, sentiment analysis research still faces challenges in NLP. Deep learning have successfully overcome the challenges of traditional machine learning models in terms of efficiency and performance. This study proposes GRU-SVM model. GRU is used because it can adaptively control dependencies, making it more efficient in memory usage, while SVM is used as it is state-of-the-art in sentiment analysis. Result of the study show that the selection of word representation techniques, the addition of feature extraction techniques, datasets, data ratios, and feature expansion are crucial in the model testing process. The GRU-SVM model achieved the best performance with an accuracy of 96.02%. Overall, the results of this study demonstrate that the GRU-SVM method is effective in analyzing sentiments in Indonesian tweets.

Keywords: *granularity, gru, sentiment analysis, social media, svm.*

1. INTRODUCTION

Sentiment analysis has seen an increase in its application in recent years, such as emotion detection [1], stock prediction [2], and tracking human habits [3]. Initially, the main methods were based on rule-based and lexicon-based approaches, but the performance of the models was disappointing. Most studies work in binary classification. However, sentiment in the real world has a double polarity that cannot be classified into positive or negative sentiment only.

Sentiment analysis has been researched at various levels of granularity, from coarse-grained to fine-grained. Zheng [4] performed an analysis of movie reviews, the process is divided into two parts, namely, classifying the review into positive and negative reviews, and then classifying the positive reviews into a finer level (fine-grained). The results provided are more in line with the emotions of the audience, but the accuracy is lower than binary classification [5]. Achieving fine-grained analysis can be done by using deep learning models. Deep learning model, specifically CNN, performs better on the multiclass classification task than machine learning [6].

Sequence models have been introduced because of their outstanding performance. Research on sentiment analysis using three neural network models shows that RNN has the highest accuracy but is very computationally expensive [7]. Dang [7] built various deep learning models such as CNN, RNN, and DNN and applied TF-IDF and Word Embedding on the dataset. They emphasized that Word Embedding is more appropriate than TF-IDF in sentiment analysis tasks. Various popular models have been compared in many studies. SVM, RNN, and GRU have also been compared in performance for sentiment analysis by the team Yongping, and the results showed that the GRU model achieved the highest accuracy compared to other models[8]. Li [9] focuses on improving the sentiment classification of restaurant reviews through the use of an attention-based bi-GRU neural network. The authors found that the attention mechanism helped to better weight the important words in each review and the bi-GRU architecture helped to capture the contextual information in both forward and backward directions. Liu [10] presents a deep learning model for sentiment analysis of e-commerce product reviews using Bert-BiGRU-Softmax architecture. The study found that the model achieved high accuracy in sentiment classification compared to traditional machine learning methods and other deep learning models such as BiLSTM and CNN. The results indicate the potential of using the Bert-

BiGRU-Softmax architecture for sentiment analysis in e-commerce product reviews.

GRU as the newest model from RNN can be a solution to the RNN model. The GRU model has gates that can control the memorization process and has a simpler architecture, which makes it faster than LSTM. Agarwal [11] presents a modified GRU model for sentiment analysis on tweets. The model was pre-processed and achieved the best accuracy of 97.7% on 35 epochs, classifying sentences as positive and negative. Boomatham [12] found that the LSTM and GRU models outperformed the SVM and MLP models. This suggests that the LSTM and GRU models are better suited for handling large amounts of data. However, it should be noted that LSTM and GRU models take longer to process. Recent research focuses on hybrid models by combining deep and machine learning architectures to optimize accuracy. The CLSTM-SVM method achieves higher accuracy than single models [13]. Zulqarnain [14] obtained good performance in classifying text against 10 target classification topics using the GRU-SVM model.

Tao [15] presents a novel neural network architecture that combines Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for intrusion detection in network traffic data. The GRU component is used to capture the sequential patterns in the network traffic data, while the SVM component is used for binary classification of intrusion or normal activity. The experimental results show that the proposed architecture outperforms existing intrusion detection methods in terms of accuracy and processing speed.

Saleh [16] investigates the use of hybrid and deep learning models for Arabic sentiment analysis. A novel ensemble stacking model is proposed that combines three pre-trained models: deep layers of CNN, hybrid CNN-LSTM, and hybrid CNN-GRU, with a meta-learner SVM. The proposed model showed the highest accuracy compared to other models, with an accuracy of 92.12%, 95.81%, and 81.4% for Main-AHS, Sub-AHS, and ASTD, respectively.

The model used is GRU-SVM as one of the state-of-the-art algorithms in sentiment analysis [17]. Text is represented by GloVe and the activation function used is softmax to determine the output of the model. SVM is placed at the last layer of the model as a classifier that classifies sentiment into fine-grained sentiment.

This research focuses on fine-grained sentiment analysis at the sentence level using the concepts of n-grams, TF-IDF, and word embedding on public policy topics in Indonesia. The scopes of this research that it only analyzes tweets in the Indonesian language, and the public policies analyzed are limited to five policies, namely politics, fuel price increase by Pertamina, petralite fuel quality, BPJS system, and the distribution of aid funds (bansos). The dataset used is taken at the time when the public policy was announced and just established, from January to November 2022 with 30,000 tweets.

The main objective of this research is to perform sentiment analysis at a high level of granularity related to public policies in Indonesia using the GRU-SVM model. Then, measure and analyze the performance of the model using accuracy as a metric.

This research is divided into four parts. In Section 2, we explain the model architecture and methodology in detail. Then we present and analyze the results in Section 3. Finally, we provide a conclusion in Section 4.

2. METHODOLOGY

2.1. Modeling Flow

This stage discusses the overall description of the system design in this research. The modeling is done through the process of data acquisition, labeling, survey, feature extraction, modeling, and evaluation. The complete flow of the sentiment analysis model-building process can be seen in Figure 1.

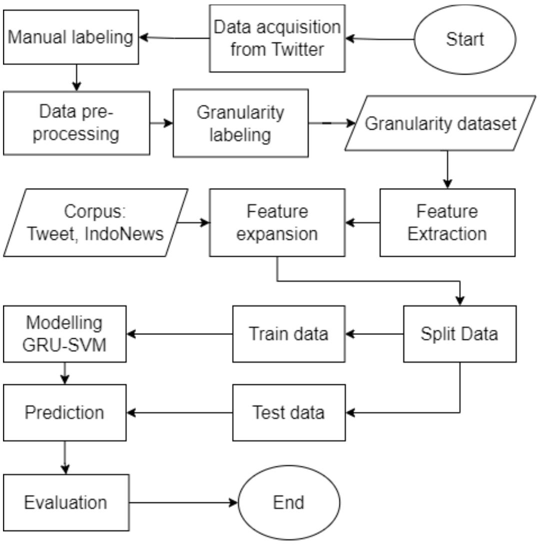

Figure 1: The flow of Sentiment Analysis Modeling

The modeling process shown in Figure 1 starts with collecting data from Twitter. Three individuals then manually label the data, and the final label is determined through majority voting. The next step is to clean and preprocess the data, resulting in a manual dataset. This dataset is then divided into three categories: negative, neutral, and positive. N-gram, TF-IDF, and manual selection techniques are used to identify unique words. After conducting a survey to establish a granularity level, a granularity corpus is created. The granularity is then labeled on the manual dataset to produce a granularity dataset. To handle the imbalance in the data, text augmentation and random deletion methods are applied. Finally, features are extracted and expanded before modeling, prediction, and evaluation can take place.

2.2. Dataset

Dataset used is from tweets on Twitter through a crawling process using Snscrape. Snscrape is a web scraping tool designed for the social media platform Twitter. It extracts data such as user profiles, hashtags, searches, tweets (single or in a thread), list posts, and trends from the platform [18]. The collected dataset is tweets related to public policy in Indonesia including politics, BPJS system (Indonesian Social Security Administration), fuel price increases by Pertamina, petralite fuel quality, and the distribution of aid funds (bansos) with a total of 30,668 tweets with a balanced distribution for each label as shown in

Table 1. Data collection was done by using several hashtags and keywords including: “politik”, “pertamina”, “petralite”, “BPJS”, and “bansos”.

Table 1. Labels Distribution of Manual Dataset

Labels	Number
Negative	10525
Neutral	9851
Positive	10292
Total	30668

Table 1 shows the distribution of the manual dataset labels. The total number of texts in the dataset is 30,668. Manual dataset has a balanced distribution of sentiment labels with an approximately equal number of negative, neutral, and positive instances. After that, the data is manually labeled into three sentiment labels: negative, neutral, and positive, as shown in Table 2. The dataset is then referred to as the manual dataset. The labeling is based on the following:

1. Positive: appreciation, satisfaction, calmness, surprise, happiness, gratitude, value.
2. Neutral: objective information.
3. Negative: sadness, neglect, fatigue, boredom, disgust, anger, frustration, annoyance.

Table 2. Manual Dataset Labeling Results

Text	Sentiment
@bread_selling korupsi bansos mah levelnya bajingan diatas bajingan, orang2 lagi kena musibah ditilep duit bantuannya.	negative
Cara Daftar Bansos BPNT 2022 Online Lewat HP di Aplikasi Cek Bansos https://t.co/bNwB89ZSoI	neutral
"Untuk masyarakat2 kurang mampu dipedesaan bansos ini pastinya sngat membantu, mereka sngt senang dpt bantuan yg langsung diterima dn dpt digunakan utk Daya Beli Masyarakat #BansosJokowi https://t.co/bYPWphWO7H"	positive

Table 2 shows the results of manual labeling of a dataset. For example, the first text is a tweet about corruption in a relief program and it is labeled as negative. The second text is a tweet about how to apply for relief program and it is labeled as neutral. The third text is a tweet about how relief program helps people in rural areas and it is labeled as positive.

In addition, this research also applies the concept of n-grams, such as 1-gram, 2-gram, 3-gram, and a combination of all three, to perform sentiment predictions. The labeled dataset is divided into three documents: negative, neutral, and positive. After applying the n-gram method, unique words are selected by performing the TF-IDF process on the documents according to the size of each n-gram. After being selected with the TF-IDF method, the unique words are also selected manually. Then, a survey is conducted to obtain the granularity level of each word according to human perception. The number of words surveyed is 300 words for each unigram, bigram, and trigram. The sample set of tweets that have gone through the n-gram, TF-IDF, and survey process is shown in Table 3.

Table 3. Granularity Dataset

Word	Negative (%)	Neutral (%)	Positive (%)	Skor
geram	68	26	5	-4
bahagia	0	5	95	5
bbm rasional	11	58	32	0
aduh malas banget	84	16	0	-4
bansos bantu masyarakat	0	11	89	4

For example, in Table 3, for the word "geram", the survey results indicate that the word "geram" has a negative score with 68% of respondents considering the word negative, 26% of respondents considering the word neutral, and only 5% of respondents considering the word positive. The score shown as -4 indicates that the word "geram" is considered negative by most respondents. This can be used to evaluate the sentiment or feeling generated by the word in a certain context. The scores in Table 3 are obtained from the conversion of the largest percentage value between negative, neutral, and positive, ranging from -5 to 5. The number of unigrams, bigrams, and trigrams consists of 900 words which are used as a new corpus for labeling datasets using granularity. Granularity labeling is done by associating the value of the word corpus granularity to each word in the text in the dataset. Label distribution results from manual labeling and granularity are shown in Table 4.

Table 4. Distribution of Labels in Dataset

Dataset labeling techniques	Negative	Neutral	Positive
Manual	10525	9851	10292
Granularity (*imbalance*)	3103	20413	7152

As seen in Table 4, the dataset from manual labeling has a balanced distribution, but the dataset generated from granularity labeling has an imbalanced distribution. Then addressed by performing text augmentation on negative and positive sentiments and removing neutral sentiments randomly.

Create an embedding from scratch from the dataset using GloVe technique to see the similarity of each word in the dataset. In this research, the embedding was also built and configured with different n-gram sizes (1-gram, and 3-gram) so that it can see the similarity of 3-grams in the corpus. An example in Table 5 displays three words that have a large similarity with the words "bbm" and "bansos rawan tikus".

Table 5. Similarity of the words "bbm" and "bansos rawan tikus"

Rank	1-gram		3-gram	
	bbm		bansos rawan tikus	
	Word	Similarity score	Word	Similarity score
1	subsidi	98.17%	sabda presiden sama	61.08%
2	masyarakat	97.84%	bengkak hapus program	60.32%
3	bantu	97.48%	tadi makan enak	57.63%

Table 5 shows the similarity between the words "bbm" (which stands for "Bahan Bakar Minyak" or fuel in Indonesian) and "bansos rawan tikus" (which roughly translates to "rat-prone aid"). The word "subsidi" (aid/subsidy in Indonesian) has the highest similarity score of 98.17% with the word "bbm." The 3-gram section shows the similarity scores between three-word phrases. The phrase "sabda presiden sama" (same as the president's statement in Indonesian) has the highest similarity score of 61.08% with the phrase "bansos rawan tikus".

2.3. Text Preprocessing

The preprocessing steps that are taken before inputting the dataset into the model are cleaning, normalization, stopword removal, stemming, split data, and tokenization.

Case folding: In this process, the tweet data is cleaned from double spaces; links; numbers; hashtags; symbols; punctuation marks; and usernames using Regex, then changing all capital letters in the tweets are lowercase.

Normalization: The normalization process works to clean up words that are not in line with linguistic rules or slang words. Normalizing text will improve and standardize the words used in tweets. The normalization process in this research uses a dictionary from the Colloquial Indonesian Lexicon [19]. It provides a list of Indonesian words and their corresponding translations or explanations in English. This lexicon can be useful for natural language processing tasks such as sentiment analysis, where it is important to have a comprehensive understanding of the language being analyzed.

Stop-word removal: The next preprocessing step is to remove meaningless words including "dari", "ke", and "adalah". In this research, the researcher uses a stopword list obtained from the Sastrawi library. Indonesian stemming library known as Sastrawi is widely used [20].

Stemming: This process breaks down words with affixes into base words. For example, the word "memakan" becomes the word "makan" after the steaming process [21].

Training-test split: The data received in this process has been cleaned, organized, and ready for modeling. In this process, the data is divided into training and test data. The training data is 90% and the test data is 10%. This division is done because the model built is a deep learning model that requires a lot of data for training rather than testing.

Tokenization: Tokenization is the process of breaking a sentence into several words to produce several tokens. For example, "digging tweet data" will produce 3 tokens, namely "digging", "data", and "tweet". We perform tokenization on tweet data using TensorFlow library [22]. In this process, sequences and padding are also done on words.

2.4. Feature Extraction

Word embedding, each word will be represented as an integer so that the text in the dataset will turn into a set of integer numbers. This extraction technique is used to simplify the process of processing text by deep learning because it can reduce the dimensions of the data [23].

TF-IDF [24] technique is used to evaluate the importance of a word in a dataset by considering the number of occurrences of the word and the number of documents containing the word as shown in Equations (1). $W_{i,j}$ is the weight for term i in document j. $tf_{i,j}$ is number of occurrences of i in j. df_i is number of documents containing i. N is total number of documents.

$$W\_(i,j) = \llbracket tf \rrbracket \_(i,j) \times lo\,g\left(\frac{N}{df_i}\right) \qquad (1)$$

N-gram is used to obtain the context of the words that appear in a sentence. This technique works by cutting the sentence according to the length of n. When n is 1 it is called unigram, for example, "Pangkal"; when n is 2 it is called bigram, for example,e "Pangkal Pinang"; when n is 3 it is called trigram, for example, "Kota Pangkal Pinang". So, the context of the word is obtained which refers to a specific place name.

2.5. Feature Expansion

Feature expansion process uses GloVe [25]. Gloves have been widely used in various applications in natural language processing tasks, such as word analogies, named entity recognition, and sentiment analysis. GloVe can create a better vector representation than the previous one. Feature expansion works by replacing weights that are zero.

The GloVe will switch weights better because it has trained with a wider corpus. GloVe will generate words that are similar to words that have a value of zero. By using the initial weights determined by GloVe model, the Embedding layer will start the training process with a better vector representation than if the initial weights were.

2.6. Modelling

The proposed model is a combination of Gated Recurrent Unit (GRU) architecture and Support Vector Machine (SVM). The deep learning model used is GRU. The GRU architecture was introduced by Cho [26]. GRU is suitable to be used when the computational load is the main priority rather than accuracy. GRU is a simpler version of LSTM, so the training process will be faster as well [27]. The architecture of the GRU model can be seen in Figure 2.

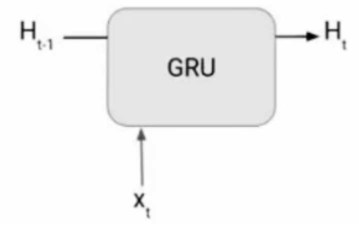

Figure 2: GRU architecture

Figure 2 illustrates the GRU model, this sequential model only has two gates, the update gate and the forget gate. Update gate helps the model determine how much past information (from the previous time step) needs to be carried forward. Forget gate is used by the model to decide how much past information should be forgotten. At each timestamp t, it requires input x_t and hidden state H_{t-1} from the previous timestamp t_1. Then it outputs a new hidden state H_t which is again passed on to the next timestamp [8], [27]. Figure **3**The architecture of the GRU-SVM model can be seen in Figure 3.

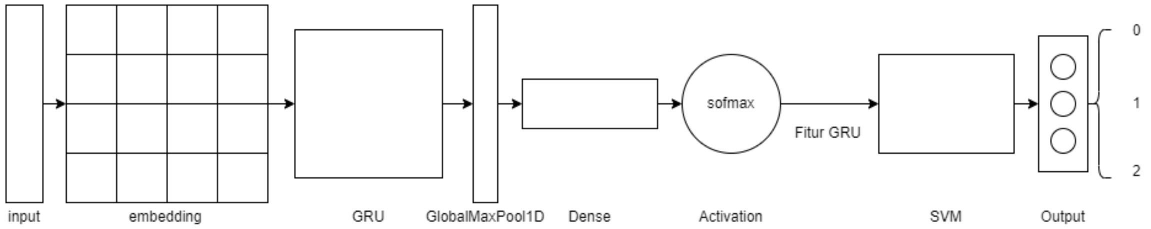

Figure 3: GRU-SVM architecture

In Figure 3, the proposed hybrid architecture combines a simple GRU with SVM with added dropout regularization, kernel regularization, GlobalMaxPool1D, and softmax activation function. Then the features of the GRU model will be classified by the structure of SVM as a classifier.

2.7. Evaluation Metric

The evaluation metric used is accuracy. Accuracy is obtained by calculating the ratio of the number of correct predictions to the total number of predictions made. This metric is chosen because the dataset used is balanced. Below is the formula for accuracy shown in Equation (2).

$$accuracy = \frac{True\ Positives + True\ Negatives + True\ Neutrals}{Total\ Observations} \quad (2)$$

From Equation (2), "True Positives" (TP) refers to the number of instances that are correctly classified as positive. "True Negatives" (TN) refers to the number of instances that are correctly classified as negative. "True Neutrals" (TN) refers to the number of instances that are correctly classified as neutral. "Total Observations" refers to the total number of instances in the dataset. Just like in binary classification, accuracy is a simple and intuitive metric that gives a good idea of how well the model is performing.

3. RESULT

This section describes the results and analysis of the GRU-SVM hybrid model to perform sentiment analysis tasks. Prediction results will be compared with test data and evaluated with an accuracy score. The following scenario used for testing can be seen in Table 6.

Table 6: Test Scenario

No	Scenario	Purpose
1	Test GRU-SVM model with manual and granularity datasets	Evaluate the model in analyzing sentiment using different datasets.
2	Test GRU-SVM model with one-hot vector, TF-IDF, and word embedding.	Determine which representation technique is most effective in predicting sentiment in text.
3	Add n-grams with a combined size of 1-gram, 2-gram, and 3-gram.	Determine the most effective n-gram size in strengthening text representations and influencing sentiment analysis results.
4	Add feature expansion techniques with corpus similarity from GloVe pre-builds with different top n sizes.	Evaluate the effect of adding this feature on the sentiment analysis results to get the optimal top similarity size

Table 6, which contains scenarios, was made to assist us in carrying out the research process so that the model can be better as more test scenarios are added. The model used to validate the developed model is a simple GRU-SVM model that only uses word embedding as feature extraction, it will also be the baseline for this research. The development focuses on feature extraction and feature expansion with GloVe. During the testing phase, we carried out a variety of tasks, including selecting representation techniques, datasets, data split size, feature extraction techniques, and feature expansion size. Testing was conducted with several datasets, including manual and granularity datasets, and also considered the selection of training and testing data ratios. Both datasets were tested by varying the split ratio to 90:10, 80:20, and 70:30 and evaluating accuracy and F1-score as metrics. First, we started by testing the model on a manual dataset, the results are shown in Table 7.

Table 7. Accuracy of Manual Dataset

Dataset	Ratio	Accuracy (%)	F1-score (%)
Manual	90:10 (*baseline*)	69.19	69.90
	80:20	68.59 (-0.87)	67.57
	70:30	68.49 (-1.01)	68.29

Table 7 shows the accuracy results of the sentiment analysis performed on the manual dataset. The dataset is split into different ratios, such as 90:10, 80:20, and 70:30, to evaluate the performance of the sentiment analysis model. The 90:10 ratio is used as the baseline to compare with the results obtained from the other ratios. The results show that the accuracy of the manual dataset ranges from 68.49% to 69.19% and the F1-score ranges from 67.57% to 69.90%. These results suggest that the 90:10 ratio provides a more robust evaluation of the model's performance. The higher accuracy and F1-score indicate that the model is better able to generalize to unseen data when using the 90:10 ratio. However, results indicate that the sentiment analysis model performs relatively poorly when using the manual dataset.

We created a granularity dataset, which had an imbalanced distribution. To address this, we performed text augmentation on the dataset. An example of text augmentation is shown in Table 8, with a sample size of three.

Table 8. Samples of Text Augmentation

Text	Augmented Text
pertamina semangat beri bahagia masyarakat	pertamina adil beri bahagia masyarakat pertamina semangat beri gembira masyarakat pertamina semangat bantu bahagia masyarakat

Augmentation results in Table 8 are obtained by replacing some words with similar words to the original word. This was done using GloVe word embedding. The text distribution of the granularity dataset was made to resemble the distribution of the manual dataset (balanced). Sentiments with negative or positive labels were augmented using NLPAUG. For example, original text is augmented by replacing certain words with synonyms to create new sentences. In the first augmented text, "pertamina adil beri bahagia masyarakat" (Pertamina is fair and brings happiness to the community). These augmented texts are used to expand the original text dataset, increasing its size and diversity. Granularity dataset distribution after the augmentation process is shown in Table 9.

Table 9. Distribution After Text Augmentation

Dataset	Negative	Neutral	Positive
Granularity	3103	20413	7152
Granularity (balanced)	10525	9851	10292

Table 9 displays a balanced granularity dataset. Negative sentiments were augmented by adding 7244 data and positive sentiments by adding 3140 data. Meanwhile, neutral sentiments were randomly deleted by 10562 data.

This research also conducted tests on the granularity dataset with varying ratios of test and training data. Table 10 shows the accuracy and f1-score results of the granularity dataset testing.

Table 10. Accuracy using Granularity Dataset

Dataset	Ratio	Accuracy (%)	F1-score (%)
Granularity	90:10	**94.97 (+37.26)**	**94.96**
	80:20	93.96 (+35.80)	93.94
	70:30	93.89 (+35.70)	93.90

Based on Table 7 and Table 10, it can be seen that the granularity dataset provides better results compared to the manual dataset in sentiment analysis. For example, at a 90:10 ratio, the granularity dataset achieved an accuracy of 94.97% and an F1-score of 94.96%, while the manual dataset only achieved an accuracy of 69.19% and an F1-score of 69.90%. The accuracy values obtained from the manual dataset are used as a baseline for comparison with the results obtained from the granularity dataset. Therefore, for further use, the granularity dataset with a 90:10 ratio is selected as the dataset to be used in this research. Next, we conducted a second test scenario related to text representation techniques. The test results are in Table 11.

Table 11. Accuracy results based on Text Representation

Word Vector	Accuracy (%)
One-hot Vector	77.83 (+12.49)
TF-IDF	69.40 (+0.30)
Word Embedding (*baseline*)	**94.97 (+37.26)**

In Table 11, word embedding representation technique provides the best results on the granularity dataset. This shows that word embedding provides better text representation and influences sentiment analysis results better compared to one-hot vector and TF-IDF. There is no significant increase between one-hot vector and TF-IDF when using the granularity dataset. On the other hand, there is a very significant increase for the granularity dataset which is the dataset that provides the best accuracy when used with word embedding.

The feature extraction technique of n-gram can increase the accuracy by 1.95% from the baseline. The size of the n-gram used is 1-gram and 3-gram. This is determined because these sizes can increase the highest accuracy among other sizes including 1-gram, 2-gram, 3-gram, 1-gram and 2-gram, 2-gram and 3-gram, and 1-gram, 2-gram, and 3-gram. Below are the test results for the n-gram size displayed in Table 12.

Table 12. Accuracy based on N-gram Size

N-gram Size	Accuracy (%)
Unigram	94.97 (+37.26)
Bigram	95.45 (+37.95)
Trigram	90.07 (+30.18)
Unigram + Bigram	95.50 (+38.03)
Unigram + Trigram	**95.87 (+38.56)**
Bigram + Trigram	87.48 (+26.43)
Unigram + Bigram + Trigram	95.60 (+38.17)

Table 12 displays the accuracy results of a sentiment analysis experiment based on different n-gram sizes. The highest accuracy was achieved with the combination of unigram and trigram, at 95.87%, which is an improvement of 38.56% compared to the baseline. After testing results were obtained using the n-gram technique, the next scenario is to add feature expansion technique with corpus similarity from pre-built GloVe. The testing results with top similarity sizes 1, 5, and 10 are displayed in Table 13.

Table 13. Accuracy After Feature Expansion

Top	Accuracy (%)		
	Tweet	IndoNews	Tweet + IndoNews
1	**96.02 (+0.02)**	82.90	95.58
5	93.92 (+35.74)	80.31	95.01
10	90.30 (+30.51)	80.12	94.98

Table 13 presents the results of the GRU-SVM testing, which utilized the GloVe technique with varying top similarities, i.e., 1, 5, and 10, The findings indicate that models exhibited comparable performance, with the model employing a top similarity of one exhibiting slightly better results. To summarize, the accuracy results are shown in Table 14.

Table 14. Accuracy of GRU-SVM model

Model	Accuracy (%)
GRU-SVM one-hot vector	60.89
GRU-SVM TFIDF	69.40
GRU-SVM Word Embedding	93.92
GRU-SVM N-gram	95.87
GRU GloVe	95.56
GRU-SVM GloVe	**96.02**

As depicted in Table 14, GRU-SVM with word embedding vector representation attained the highest accuracy of 93.92% when compared to one-hot vector and TF-IDF. Furthermore, the implementation of the N-gram feature extraction technique and feature expansion with GloVe also led to improved accuracy, with values of 95.87% and 96.02% respectively.

4. DISCUSSION

In this paper, we proposed a hybrid model of GRU and SVM for sentiment analysis. The results of this research showed that the vector representation technique, feature extraction, dataset, training and test data ratio, and feature expansion have a significant impact on the performance of the GRU-SVM model. The results obtained in this study are compared with the results of a previous study on sentiment analysis using [11], as well as a recent study on [10]. One of the key factors contributing to the high accuracy of our model is the use of GRU, a type of Recurrent Neural Network (RNN) that is capable of capturing long-term dependencies in sequential data. The gating mechanism in GRU allows the model to selectively preserve and update information, leading to better performance in capturing the sentiment of social media posts. Additionally, the integration of SVM as a classification algorithm helped to further enhance the accuracy of the model. Concerning the dataset, the results from the granularity dataset showed better results than those from the manual dataset. The appropriate selection of training and test data ratio can affect model performance. The 90:10 ratio provided the highest accuracy score of 94.97%. In this sentiment analysis case, word embedding representation is more appropriate than one-hot vector and TF-IDF. Word embedding is a more robust technique compared to one-hot vector and TF-IDF. The addition of feature extraction techniques like n-gram can increase the accuracy of the model by 1.95% while adding feature expansion increases accuracy by 2.10%. The results show that the proposed GRU-SVM model outperforms both the Modified GRU and the Bert-BiGRU-Softmax models in terms of accuracy. These results demonstrate the effectiveness of the proposed model in performing fine-grained sentiment analysis on social media text data. A line chart of the increase in each scenario compared to the baseline is depicted in Figure 4.

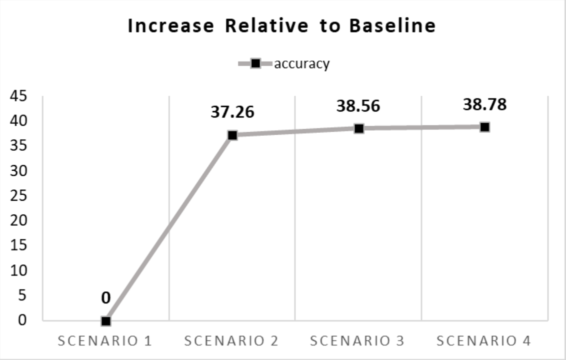

Figure 4: Increase in Accuracy to Baseline

Figure 4 shows that each scenario can enhance the performance of the model. Scenario 1 is the selection of datasets and the size of the train test split.

Scenario 2 is GRU-SVM combined with word embedding. Scenario 3 involves the addition of n-grams, which increases the accuracy by 0.02 compared to the baseline. Lastly, the features expansion with GloVe is applied.

The results of the proposed GRU-SVM model are compared with those of the Modified GRU model. The Modified GRU model uses a GRU network to extract features from the input text and a multi-layer perceptron to perform sentiment classification. The results show that the GRU-SVM model outperforms the Modified GRU model in terms of accuracy. This suggests that the combination of GRU and SVM provides a more effective solution for sentiment analysis. The SVM component of our model can effectively classify the sentiment based on the features learned by the GRU component. This combination of neural and non-neural techniques results in a more robust and accurate sentiment analysis model.

The Bert-BiGRU-Softmax model uses a combination of the Bidirectional GRU (BiGRU) and the Softmax activation function to perform sentiment classification. The results show that the GRU-SVM model outperforms the Bert-BiGRU-Softmax model in terms of accuracy.

Despite the promising results obtained in this study, some limitations should be considered. Dataset used in this study is limited to social media text, which may not generalize to other text domains.

One of the key strengths of the proposed model lies in its ability to accurately capture the nuances in sentiment, as demonstrated by its high accuracy. The granularity is crucial for effectively analyzing sentiments in social media, where opinions and emotions can be complex and nuanced.

5. CONCLUSION

This study discusses the performance evaluation of the Gated Recurrent Unit-Support Vector Machine (GRU-SVM) model for sentiment analysis on Indonesian tweets. We apply sentiment analysis with a level of granularity. The GRU-SVM model was proposed because it is a deep learning method that can solve problems in the performance and efficiency of traditional machine learning models. This research indicates that the GRU-SVM hybrid model can effectively perform fine-grained sentiment analysis on Indonesian language tweets. The researcher conducted several testing scenarios. Selection of representation techniques including one-hot vector, TF-IDF, and word embedding to transform the input before entering the model. These representation techniques were chosen on two datasets, manual and granularity datasets. The granularity dataset provides better results than the manual dataset. Next, the addition of n-gram feature extraction techniques can increase the accuracy of the GRU-SVM model combined with word embedding. This study also shows that the choice of the proper feature extraction technique is crucial in the testing process of the model. The addition of the feature expansion technique with GloVe achieved the best accuracy. The highest accuracy value is obtained when GRU-SVM is combined with the Top 1 feature using a tweet corpus of 96.02% with an increase in accuracy to baseline of 38.78%. Combining deep learning models with SVM (hybrid) produces better accuracy than using a single model. Thus, the GRU-SVM hybrid model can be considered as a method that can be used in sentiment analysis on Indonesian-language tweets. Therefore, this GRU-SVM hybrid model can be considered as one of the methods that can be used in sentiment analysis of Indonesian language tweets. In addition, this study also shows that sentiment analysis with granularity can provide more accurate results. For further research, it can be tried to use other feature extraction methods such as BERT or ELMO or add layers to the model such as attention layer to increase the accuracy in sentiment analysis of Indonesian language tweets. Furthermore, the study can also evaluate the model on a larger and more diverse dataset and evaluate the model in different domains such as sentiment analysis in other texts such as news articles or product reviews.

REFERENCES

[1] Z. Chen *et al.*, “Emoji-powered Sentiment and Emotion Detection from Software Developers’ Communication Data,” *ACM Transactions on Software Engineering and Methodology*, vol. 30, no. 2, Mar. 2021, doi: 10.1145/3424308.

[2] T. W. Sagala, M. S. Saputri, R. Mahendra, and I. Budi, “Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis,” in *ACM International Conference Proceeding Series*, Jan. 2020, pp. 123–127. doi: 10.1145/3379310.3381045.

[3] H. Jang, E. Rempel, I. Roe, P. Adu, G. Carenini, and N. Z. Janjua, “Tracking Public Attitudes Toward COVID-19 Vaccination on Tweets in Canada: Using Aspect-Based Sentiment Analysis,” *J Med Internet Res*, vol. 24, no. 3, p. e35016, Mar. 2022, doi: 10.2196/35016.

[4] J. Zheng, L. Zheng, and L. Yang, “Research and Analysis in Fine-grained Sentiment of Film Reviews Based on Deep Learning,” in *Journal of Physics: Conference Series*, Jul. 2019, vol. 1237, no. 2. doi: 10.1088/1742-6596/1237/2/022152.

[5] M. Munikar, S. Shakya, and A. Shrestha, “Fine-grained Sentiment Classification using BERT,” *arXiv preprint*, Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.03474

[6] A. Soufan, “Deep learning for sentiment analysis of Arabic text,” in *ACM International Conference Proceeding Series*,

Mar. 2019. doi: 10.1145/3333165.3333185.

[7] N. C. Dang, M. N. Moreno-García, and F. de la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics (Switzerland)*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/electronics9030483.

[8] Y. Xing and C. Xiao, "A GRU Model for Aspect Level Sentiment Analysis," in *Journal of Physics: Conference Series*, Sep. 2019, vol. 1302, no. 3. doi: 10.1088/1742-6596/1302/3/032042.

[9] L. Li, L. Yang, and Y. Zeng, "Improving sentiment classification of restaurant reviews with attention-based bi-gru neural network," *Symmetry (Basel)*, vol. 13, no. 8, Aug. 2021, doi: 10.3390/sym13081517.

[10] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, Nov. 2020, doi: 10.3934/MBE.2020398.

[11] A. Agarwal, P. Dey, and S. Kumar, "Sentiment Analysis using Modified GRU," in *ACM International Conference Proceeding Series*, Aug. 2022, pp. 356–361. doi: 10.1145/3549206.3549270.

[12] S. Boonmatham and P. Meesad, "Stock Price Analysis with Natural Language Processing and Machine Learning," in *ACM International Conference Proceeding Series*, Jul. 2020. doi: 10.1145/3406601.3406652.

[13] C. N. Dang, M. N. Moreno-García, and F. de La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9986920.

[14] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, "Text classification based on gated recurrent unit combines with support vector machine," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3734–3742, 2020, doi: 10.11591/ijece.v10i4.pp3734-3742.

[15] A. F. M. Agarap, "A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data," in *ACM International Conference Proceeding Series*, Feb. 2018, pp. 26–30. doi: 10.1145/3195106.3195117.

[16] H. Saleh, S. Mostafa, L. A. Gabralla, A. O. Aseeri, and S. El-Sappagh, "Enhanced Arabic Sentiment Analysis Using a Novel Stacking Ensemble of Hybrid and Deep Learning Models," *Applied Sciences (Switzerland)*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12188967.

[17] G. Balikas, S. Moura, and M. R. Amini, "Multitask Learning for Fine-Grained Twitter Sentiment Analysis," in *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2017, pp. 1005–1008. doi: 10.1145/3077136.3080702.

[18] JustAnotherArchivist, "snscrape: A social networking service scraper in Python," *2018.*

[19] M. Dong, Universitas Telkom, Chinese and Oriental Languages Information Processing Society, Institute of Electrical and Electronics Engineers. Indonesia Section. Computer Society Chapter, and Institute of Electrical and Electronics Engineers, "Proceedings of the 2018 International Conference on Asian Language Processing (IALP) : 15-17 November 2018, Telkom University, Bandung, Indonesia".

[20] N. Yusliani, R. Primartha, and M. D. Marieska, "Multiprocessing Stemming: A Case Study of Indonesian Stemming," 2019. [Online]. Available: https://github.com/sastrawi

[21] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia."

[22] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." [Online]. Available: www.tensorflow.org.

[23] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey."

[24] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl Inf Syst*, vol. 60, no. 2, pp. 617–663, Aug. 2019, doi: 10.1007/s10115-018-1236-4.

[25] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation." [Online]. Available: http://nlp.

[26] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.1078

[27] Z. Penghua and Z. Dingyi, "Bidirectional-GRU based on attention mechanism for aspect-level Sentiment Analysis," in *ACM International Conference Proceeding Series*, 2019, vol. Part F148150, pp. 86–90. doi: 10.1145/3318299.3318368.