Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 9, No. 1, March 2023, pp. 119-131
ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i1.25813

Social Media Sentiment Analysis Using Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU)

Ahmad Zahri Ruhban Adam, Erwin Budi Setiawan
Telkom University, Jl. Terusan Buah Batu, Bandung 40257, Indonesia

ARTICLE INFO

Article history:

Received February 08, 2023
Revised February 28, 2023
Published March 01, 2023

Keywords:

Convolutional neural network;
Gated recurrent unit;
Sentiment analysis;
Feature expansion;
Twitter

ABSTRACT

The advancing technologies are aimed to maximize human performance. One of the great developments in technology is social media. The social media used in this study is Twitter because commonly people in Indonesia give their opinions to the public through tweets. The opinions given are very diverse, where they write positive, negative, and neutral opinions in a large collection of data. Deep learning can be used to automate the process that understands, obtains, and processes the expression of data in the form of text to obtain information from sentiment categories contained in the data. The purpose of this study is to analyze the sentiments of the opinions given by the public in Bahasa Indonesia using deep learning methods and variations in scenarios. To conduct sentiment analysis, tweets are collected by crawling the data. Tweets are then labeled positive, negative, and neutral and then represented as 1, -1, and 0. The method used to classify tweet sentiment is the Convolutional Neural Network (CNN) and Gated Recurrent Unit method (GRU). Research stages include feature selection, feature expansion, preprocessing, and balancing with SMOTE. The highest accuracy value obtained on the CNN-GRU model with an accuracy value of 97.77% value. Based on these tests, it can be concluded that sentiment analysis research on Twitter social media using the combination of Convolutional Neural Network and Gated Recurrent Unit methods can produce fairly high accuracy, and feature expansion testing of the deep learning model paired with SMOTE can provide a significant increase in accuracy values.

Corresponding Author:

Erwin Budi Setiawan, Telkom University, Jl. Terusan Buah Batu, Bandung 40257, Indonesia
Email: erwinbudisetiawan@telkomuniversity.ac.id

1. INTRODUCTION

The advancing technologies are aimed to maximize human performance. One of the great developments in technology is social media. Social media can be a reflection of a user's personality because information spread on social media has a major impact on its users. Social media that has a lot of users in Indonesia is Twitter. Twitter users can ‘tweet’ to share various information in the form of text, videos, and images. One can find a large number of diverse opinions from the public through twitter in Indonesia which can be positive, neutral, and negative. Those various opinions needs a system to classify the sentiment.

Classification is the development of machine learning which is included in the category of supervised learning methods [1]–[3]. Classification is a process to predict the class of the given data. Class is defined as a category or label/target. Various methods can be used to classify sentiment data [4]–[7]. Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) are two of them [8]–[14]. Convolutional Neural Network (CNN) is a deep learning algorithm that can receive input in the form of images, distinguish one image from another, and determine various aspects and objects in an image that can be 'learned' by a machine. Gated Recurrent Unit (GRU) is a gating mechanism of Recurrent neural networks. GRU is similar to long short-term memory (LSTM) but has fewer parameters than LSTM. GRU aims to enable each recurrent unit to capture dependencies on different but adaptive time scales.

This research was conducted to implement sentiment analysis to see other people opinions about certain topics. Sentiment analysis is a technique that can be used to identify sentiments expressed in a text and classify these sentiments into positive, negative or neutral sentiment categories. Sentiment analysis also means an automatic process that understands, obtains and processes the expression of data in the form of text to obtain information from sentiment categories contained in the data [15]–[19].

Many studies have been conducted for sentiment analysis, the methods used are also varied, traditional machine learning methods and deep learning [20]–[22]. However, sentiment analysis research using the CNN-GRU method is still relatively small. Research by Candradinata, et al conducted a Sentiment Analysis on Twitter regarding Online Store Services using the Naïve Bayes Method. The purpose of this research is to find out the user's sentiment towards the company, so that the user knows whether the system used is good for the user or not. The data used is data from Twitter. Data is divided into 3 classes: positive, negative and neutral. The results of this study are the highest average performance, namely accuracy of 66.64%, precision of 67.13%, and recall of 68.44% [23].

Alkahfi and Chiuloto conducted a study entitled "Application of the Gated Recurrent Unit Model During the Covid-19 Pandemic in Predicting Gold Prices Using the Mean Square Error Measurement Model". The purpose of this research is to forecast gold prices to make it easier for the public to see the market value of gold for the next few months. This study uses the Gated Recurrent Unit (GRU) method for gold price forecasting and then uses an error level measurement, namely the Mean Square Error (MSE) which functions to check the error value in gold forecasting. The results obtained are the MSE error rate of 0.111 and RMSE of 0.334 and R-squared of 0.5 [24].

Research related to the Classification of Cat and Dog Sounds Using LSTM-GRU and ANN-BP has purpose to do machine learning by giving the sound of a dog or cat using the LSTM-GRU and ANN-BP methods and then the machine will determine whether it is a cat or dog sound. The data used in this study comes from the Kaggle Repository: https://www.kaggle.com/mmoreaux/audio-cats-and-dogs. The data contains the sounds of cats and dogs uploaded by Marc Moreaux. With data consisting of 277 files in (.wav) format, the accuracy obtained is 92% with a precision of 0.91 and a recall of 0.91 [25].

Based on the research above, this research is trying to develop a system for classifying sentiment analysis that is built by expanding the fastText feature on the Indonesian Tweet dataset on Twitter using the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) methods. Those methods are chosen because the CNN model has the advantage of automatically extracting important features from each data, besides that the CNN method is also more efficient than other neural network methods, especially for memory and complexity [8]. while GRU is for making each recurrent unit to be able to capture dependencies in different time scales adaptively. As an analogy, humans do not need to use all the information in the past to make decisions now [12]. As far as the researchers know, there has been no research in Indonesia that has conducted sentiment analysis using 2 CNN and GRU deep learning models and combining 2 methods into CNN-GRU and GRU-CNN. Therefore, this research was conducted to determine the effectiveness of a sentiment analysis system that uses 2 deep learning methods and combines them, in processing data that contains opinion sentences compared to using conventional sentiment analysis procedures. The purpose of this study is to implement a tweet classification system using the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) methods with several scenarios such as split data selection, feature extraction, and feature selection. After that, feature expansion and the CNN-GRU method are tested following with GRU-CNN using fastText corpus and data balancing using SMOTE to prevent data overfit. This stages are conducted to get results with high accuracy and efficiency to make predictions. This research uses two types of data, namely manual labeling data and granularity labeling as a comparison of system performance in carrying out tasks.

Further discussion in this paper will contain the following. Section 2 contains a description of the research method regarding Social Media Sentiment Analysis using Convolutional Neural Network (CNN) dan Gated Recurrent Unit (GRU). Section 3 contains results and discussion and followed by the conclusions drawn in Section 4.

2. METHODS

2.1. Research Design

The following picture is an overview or Flowchart of designing Sentiment Analysis based on Tweets using the Convolutional Neural Network and Gated Recurrent Unit methods. Flowchart is given in Fig. 1.

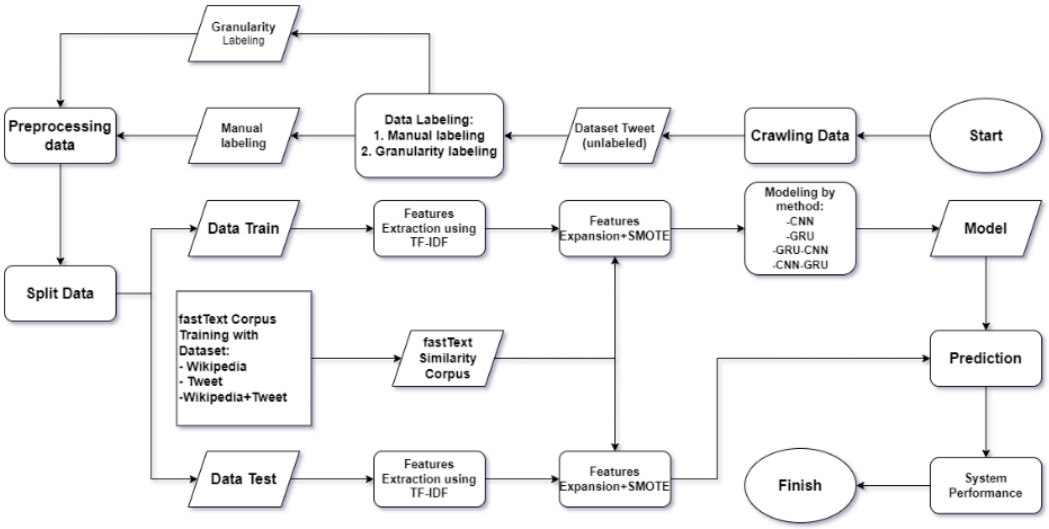

Fig. 1. Research Stages of Sentiment Analysis

2.2. Crawling Data

The dataset used for this research was obtained through the results of crawling data from Twitter. Crawling is retrieving data from data sources that will be used as a dataset. In this research, the data source comes from Twitter in Bahasa Indonesia. Data is taken using the API provided by Twitter. The initial stage to get access to Twitter data is to register an account as a developer and describe why you want to access the API key from Twitter.

Crawling works by connecting the Twitter API into a python script, then getting an output dataset in the form of .CSV or .XLS which contains user IDs and tweets. Tweet topics used in this dataset are "Politik", "Pertamina", "BPJS", "Pertalite", and "Bansos" as shown in Table 1.

Table 1. Tweet's Topics

Topic	Total
Politik	8,313
Pertamina	8,492
BPJS	10,135
Pertalite	583
Bansos	10,000

2.3. Data Labelling

2.3.1. Manual Labelling

After the data from Twitter has been obtained, then the data is labeled manually into three different labels, namely positive, negative and neutral. Labeling involves three annotators with the principle of majority voting. Table 2 bellow is example of manual data labeling and distribution labels and Table 3 contains the distribution of the labels.

Table 2. Example of Data Labelling

Tweet	Label 1	Label 2	Label 3	Majority Label
hehehehe ahok sdh tenang di pertamina move on dnk 😊 apakah blm ada bahan buat membully pj gub 😝 ayo bisa yuk kamu bisaaah	Negative	Positive	Negative	Negative
satgas tni manunggal membangun desa tmmd sengkuyung tahap iii tahun 2022 kodim 0707/wonosobo secara bergotong royong bersama warga masyarakat desa maron membuat senderan sebagai tanggul menahan tanah agar tidak mudah longsor	Positive	Positive	Positive	Positive

Table 3. Distribution of Data Labels

Label	Sum	Percentage
Positive	10417	33.80%
Negative	10804	35.06%
Neutral	9590	31.12%
Total	30811	100%

2.3.2. Labelling by System

In this labeling by system, the researcher created a corpus containing 900 lists of words and sentences of words of sentiment, after that a survey was conducted to give positive, negative, and neutral values for each word and sentence in the survey. In Fig. 2 you can see the design of the labeling system.

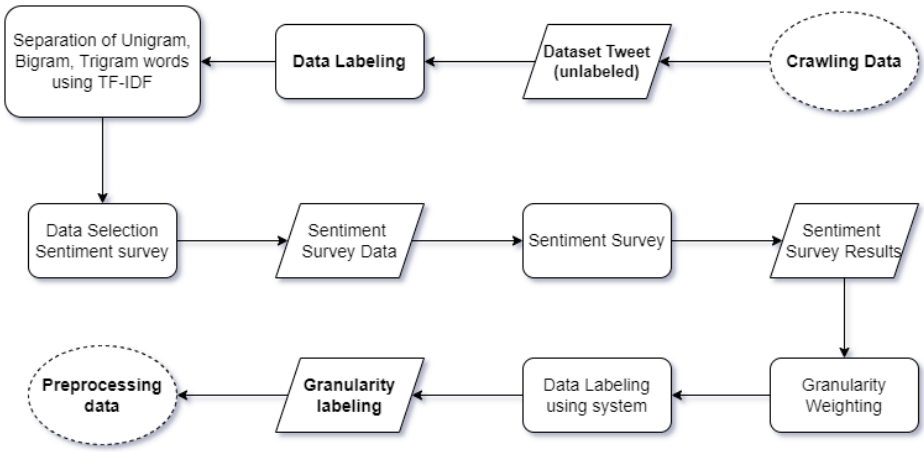

Fig. 2. Granularity Labelling

Table 4 shows the results of several examples of words and sentences used in the survey. In Table 5 you can see the results of the distribution of the word sentiment survey.

Table 4. Example of Survey Sentiment

No	Word	Negative (%)	Neutral (%)	Positive (%)	Granularity
1	Akrab	0%	3%	97%	5
2	Adiktif	63%	30%	7%	-4
3	Usang	53%	33%	13%	-3
4	Bakti sosial	3%	17%	80%	5
5	Azab mati	87%	13%	0%	-5
6	Gotong royong	0%	10%	90%	5
7	Kasihan rakyat	37%	37%	27%	-2
8	Apbd bantu sosial	7%	43%	50%	3
9	Subsidi bahan bakar	3%	53%	43%	0
10	Terima kasih presiden	3%	17%	80%	4

Table 5. Distribution of Granularity Labelling

Label	Sum	Percentage
Positif	5864	19.03%
Negatif	4244	13.77%
Netral	20703	67.19%
Total	30811	100%

2.4. Pre-processing

The next process is data preprocessing. The data on Tweets contains noise, moreover, the data on tweets is irregular and quite complicated. Thus, preprocessing is needed to remove all meaningless and necessary characters so that only the important words are left. The results of preprocessing are better in terms of the occurrence of words that are significant and less meaningful. Preprocessing is an important step to do because

if there is a lot of redundant information and words that are less relevant, then the system classification phase will be more difficult [26]. The processes include:

1. Data cleaning is a process that includes the removal of numbers, characters other than alphabetic and numeric, and non-ascii characters.
2. Case folding is a process that changes the words in a Tweet to lowercase. This process is useful if there are words written with different capitalization.
3. Tokenizing is the process of separating or cutting tweets into the words that make them up. This process includes removing punctuation marks, characters other than the alphabet. This is done to prevent noise in the next.
4. Stopwords removal is the process of removing non-topic words that are not important, in this case words that are included in the stoplist, one of which is connecting words such as "which", "and", "then", etc.
5. Stemming is the process of removing a word and turning it into a base word. It works by removing a prefix or suffix from a word

2.5. Feature Extraction

Feature extraction is the process of extracting or converting a document into a text format into features that can be easily processed by machine learning classification techniques. Feature extraction is one of the most important techniques in data mining and calculating feature values in documents [26]. In this research, TF-IDF is used for weighting and feature extraction. The term frequency (TF) is normalized by the inverse document frequency (IDF). This method is often used primarily for information retrieval. The TF-IDF calculation is defined as follows [6].

$$tfidf(k,T) = tf(k,T) \times idf(k) \quad (1)$$

$$idf(k) = \log \frac{1+n}{1+df(k)} + 1 \quad (2)$$

Where $tf(k, T)$ is the number of words to search for in a document, n is the total document tweet, and $df(k)$ is the number of documents containing the word k.

2.6. Feature Expansion

The next stage is feature expansion which is used to overcome the non-appearance of words by replacing similar words with groups of features that have a high degree of similarity. Feature expansion is conducted using fastText [26]–[31]. In its process, FastText makes use of sub-words that use a skip-gram model vector with n-gram characters developed by AI research team of Facebook [30][32]. First, fastText is trained using the Wikipedia dataset to produce a fastText similarity corpus that can get the similarity of a word. As an example, it can be seen in Table 6 the list of words that have similarities to the word "pertalite" which has been sorted according to its similarity value.

Table 6. Similarity Words from "pertalite"

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
Pertalite	dexlite	pertamax	Dexlite	DEXlite	Pertamax	premium	beroktan	oktannya	BBM-nya

2.7. SMOTE

In this study, the distribution of data on Granularity Labeling is unbalanced, to prevent overfitting of the model, it is necessary to apply a balancing Synthetic Minority Oversampling Technique (SMOTE). The way SMOTE works is by adding artificial data to the minority class by interpolating the original data, so that the resulting artificial data varies. Several studies related to sentiment analysis have used smote for imbalanced class cases [33]–[39].

2.8. Classification Algorithm

There are two classification models used for this research: Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU). As illustrated in Fig. 1, the algorithm is used to create a topic classification of model tweets. The following is an explanation of each algorithm.

2.8.1. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a deep learning algorithm which is a multilayer neural network using perceptons for supervised learning and to analyze data. It can receive input in the form of images, distinguish one image from another, and determine various aspects and objects in the image that can be 'learned' by the machine [40]. Convolutional Neural Network architecture consists of 3 layers, there are input, hidden and output. Input serves to input a number of data with neurons in the feature. Hidden layer is a layer that is hidden according to the model and data size. And output is a function to change the output of each class.

2.8.2. Gated Recurrent Unit (GRU)

Gated Recurrent Unit is a variant of Long Short-Term Memory that is quite popular [41][42]. The advantage of this method is that GRU has simpler computations than LSTM, but has equal accuracy and still Update gate functions to assist the model in determining what information needs to be forwarded to the future from past information (previous time step). Meanwhile, the reset gate function is for the model to decide how much information from the past should be forgotten, which is quite effective in avoiding the gradient from disappearing. Gru has two gates, namely the reset gate and the update gate.

2.9. System Performance

The results of the classification need to be evaluated to determine the performance of the tested model. One of the performance evaluations which can be calculated using the confusion matrix is accuracy. The results of the performance evaluation can be obtained using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In the system performance section, the accuracy value is measured, and the accuracy value in this study is obtained using the Confusion Matrix. There are four terms in the Confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Table 7 is an overview table of the Confusion Matrix.

Table 7. Confusion Matrix

Class	Prediction	
	Positive	**Negative**
Positive	TP (True Positive)	FN (False Negative)
Positive	FP (False Negative	TN (True Negative)

To facilitate the comparison of performance in this study with related research that has been discussed previously, the performance measurement used in this study is **accuracy**.

3. RESULTS AND DISCUSSION

In this research, there are several stages of testing scenarios to find the best performance results. This experiment was carried out to find out whether there is an increase in terms of accuracy when several scenarios are carried out in the hope that the accuracy value can increase at each stage of the test. Testing was carried out using two types of datasets, namely Manual labeling (Manual) and Granularity labeling (System).

3.1. Result

3.1.1. Baseline and Feature Selection

The first test was conducted to determine the baseline by looking for the best Baseline Unigram, Bigram, Trigram, or Allgram proportion scenarios and determining the selection of the number of TF-IDF features. Baseline The scenario used for split data is with the proportion of 80:20. Following are the results of testing the Baseline scenario displayed in Table 8 and Table 9. With the results obtained in Table 8, the proportion of split data to be used for the next testing stage is 80:20.

With these results, the Baseline scenario used for feature selection on TF-IDF is to use Allgram and dataset labeling by system. Next, we tested the use of Max Feature on TF-IDF. From previous tests, the best Baseline scenario from Allgram was obtained using the labeling by system dataset. Then with this scenario, the Max feature selection test was carried out on the TF-IDF and displayed in Table 10.

Table 8. Results of Data Split

Unigram		Accuracy (%)	
		CNN	***GRU***
70:30	Manual	69.84	69.36
	System	88.21	87.13
80:20	Manual	**70.55**	**70.17**
	System	**89.32**	**88.64**
90:10	Manual	71.31	70.76
	System	89.29	88.61

Table 9. Results of Baseline Scenario

Condition		Accuracy (%)	
		CNN	**GRU**
Unigram	Manual	70.01	69.99
	System	89.32	88.64
Bigram	Manual	63.02	63.11
	System	79.79	80.14
Trigram	Manual	54.86	54.79
	System	74.13	74.28
Allgram	Manual	69.36	70.06
	System	**91.14 (+2.03)**	**90.73 (+2.35)**

Table 10. Results for Featrures Selection

TF-IDF	Accuracy (%)	
	CNN	**GRU**
2,000 max features	89.30 (-2.01)	89.61 (-1.23)
4,000 max features	90.70 (-0.48)	90.16 (-0.62)
6,000 max features	91.06 (-0.08)	90.54 (-0.20)
8,000 max features	91.12 (-0.02)	90.84 (+0.12)
10,000 max features	**91.93 (+0.86)**	**90.88(+0.16)**
15,000 max features	90.73 (-0.44)	90.07 (-0.72)

The results of these tests show that there is no significant increase in accuracy. It can be seen that for every max feature tested, starting from 2,000 max features up to 10,000, there is an accuracy increase of about 1%, but in testing 15,000 features there is a decrease in accuracy and also the F1 score. With this, it can be concluded that the selection of TF-IDF features does not guarantee a significant increase in accuracy.

Based on these data, the best scenario model is obtained by 10,000 max features. That way, the baseline obtained is the 80:20 data split scenario and TF-IDF with a max feature of 10,000 words. Subsequent tests test feature expansion against the baseline

3.1.2. Feature Expansion and Methods Combination

This test is carried out to find the best model from the baseline which is combined using feature expansion with fastText. Feature expansion is carried out on the top 1, 5, and 10 most similar words in the fastText corpus. It can be seen in Table 11 that the increase in accuracy value is found in almost every classifier model. CNN gets an increase in accuracy value of (+0.12%) at baseline + top5, GRU gets an increase in accuracy value of (+0.03%) at baseline + top1, and GRU-CNN gets an increase in accuracy value of (+0.57%).

Table 11. Results of Feature Expansion with Wikipedia Corpus

Model	Accuracy (%)			
	CNN	**GRU**	**GRU-CNN**	**CNN-GRU**
Baseline	91.14	90.73	90.57	91.20
Baseline+Top1	90.83 (-0.34)	**90.78 (+0.03)**	**91.09 (+0.57)**	91.15 (-0.05)
Baseline+Top5	**91.25 (+0.12)**	90.32 (-0.45)	90.55 (-0.02)	91.18 (-0.02)
Baseline+Top10	90.76 (-0.41)	90.16 (-0.62)	90.54 (-0.03)	90.97 (-0.25)

The feature expansion with the similarity corpus built, with the pre-trained fastText model indo wikipedia, gets an increase in accuracy values in almost all classifier models, an increase in accuracy is found in each classifier model, CNN gets an increase in accuracy values of (+0.12%) at baseline+top5 , GRU got an increase in accuracy value of (+0.03%) at baseline + top1, and GRU-CNN got an increase in accuracy value of (+0.57%) as shown in Table 11.

Expansion of features built with corpus similarity to the Tweet dataset tends to decrease in accuracy for all classifier models when the top-n features increase, as shown in Table 12. The greater the top-n value to the baseline, the greater the decrease in accuracy.

Table 12. Results of Feature Expansion With Tweet Corpus

Model	Accuracy (%)			
	CNN	**GRU**	**GRU-CNN**	**CNN-GRU**
Baseline	91.14	90.73	90.57	91.20
Baseline+Top1	90.67 (-0.51)	90.18 (-0.6)	89.94(-0.69)	91.02 (-0.19)
Baseline+Top5	90.05 (-1.19)	89.16 (-1.73)	89.30 (-1.40)	89.94 (-1.38)
Baseline+Top10	89.01 (-2.33)	88.48 (-2.47)	88.35 (-2.45)	88.98 (-1.33)

Table 13. Results of Feature Expansion With Tweet+Wikipedia Corpus

Model	Accuracy (%)			
	CNN	**GRU**	**GRU-CNN**	**CNN-GRU**
Baseline	91.14	90.73	90.57	91.20
Baseline+Top1	91.04 (-0.10)	90.62 (-0.12)	**90.58 (+0.01)**	**91.38 (+0.19)**
Baseline+Top5	91.01 (-0.14)	90.31 (-0.46)	90.50 (-0.07)	91.11 (-0.09)
Baseline+Top10	90.73(-0.44)	90.69 (-0.04)	90.26 (-0.34)	**91.27 (+0.07)**

The results of testing the feature expansion built with the Tweet+wikipedia corpus similarity have an increase in the accuracy value, although it is not significant. It can be seen in Table 13 that the GRU-CNN classifier model for Baseline+Top1 has an increase in accuracy value of (+0.01%), the CNN-GRU classifier model for Baseline+Top1 has an increase in accuracy value of (+0.19%), and the CNN-GRU model for baseline+top10 with an increase in accuracy value of (+0.07%).

3.1.3. Feature Expansion with SMOTE

This test is carried out to find the best model from the baseline which is combined using feature expansion with fastText coupled with balancing data using smote. Feature expansion is carried out on the top 1, 5, and 10 most similar words in the fastText corpus. It can be seen in Table 14 that the accuracy value for feature expansion with corpus wikipedia+smote increases significantly in each classifier model. the highest result obtained by the CNN-GRU model is against Baseline+Top1 with an accuracy value of 97.77% (+7.12%) of the Baseline value. the highest result obtained by the GRU-CNN model is against Baseline+Top1 with an accuracy value of 95.63% (+5.59%) of the Baseline value. the highest result obtained by the GRU model is against Baseline+Top5 with an accuracy value of 96.03% (+5.84%) of the Baseline value. The highest results obtained by the CNN model are Baseline + Top 5 with an accuracy value of 97.00% (+ 6.43%) of the Baseline value.

Table 14. Results of Feature Expansion+SMOTE with Wikipedia Corpus

Model	Accuracy (%)			
	CNN	**GRU**	**GRU-CNN**	**CNN-GRU**
Baseline	**91.14**	**90.73**	**90.57**	**91.20**
Baseline+Top1	96.80 (+6.21)	95.35 (+5.09)	**95.63 (+5.59)**	**97.77 (+7.12)**
Baseline+Top5	**97.00 (+6.43)**	**96.03 (+5.84)**	95.38 (+5.31)	97.46 (+6.86)
Baseline+Top10	96.70 (+6.10)	95.38 (+5.12)	95.02 (+4.91)	97.62 (+7.04)

The results of testing the feature expansion+smote with the tweet corpus resulted in a significant increase in accuracy values. The highest result obtained by the CNN model is Baseline+Top1 with an accuracy value of 96.61% (+6.02%) of the Baseline value. The highest result obtained by the GRU model is against

Baseline+Top1 with an accuracy value of 95.24% (+4.97%) of the Baseline value. The highest result obtained by the GRU-CNN model is against Baseline+Top5 with an accuracy value of 95.46% (+5.40%) of the Baseline value. And the highest result obtained by the CNN-GRU model is against Baseline+Top1 with an accuracy value of 97.44% (+6.84%) of the Baseline value as shown in Table 15.

Table 15. Results of Feature Expansion+SMOTE with Tweet Corpus

Model	Accuracy (%)			
	CNN	GRU	GRU-CNN	CNN-GRU
Baseline	**91.14**	**90.73**	**90.57**	**91.20**
Baseline+Top1	**96.61 (+6.02)**	**95.24 (+4.97)**	95.39 (+5.32)	**97.44 (+6.84)**
Baseline+Top5	96.53 (+5.91)	94.70 (+4.37)	**95.46 (+5.40)**	96.95 (+6.30)
Baseline+Top10	96.33 (+5.69)	94.58 (+4.24)	95.17 (+5.08)	96.93 (+6.28)

The results of testing the feature+smote expansion with the tweet+wikipedia corpus also resulted in a significant increase in accuracy values as shown in Table 16. The highest result obtained by the CNN model is the Baseline+Top5 with an accuracy value of 96.96% (+6.39%) of the Baseline value. The highest result obtained by the GRU model is against Baseline+Top10 with an accuracy value of 95.55% (+5.31%) of the Baseline value. The highest result obtained by the GRU-CNN model is against Baseline+Top10 with an accuracy value of 95.69% (+5.65%) of the Baseline value. And the highest result obtained by the CNN-GRU model is against Baseline+Top5 with an accuracy value of 97.58% (+6.99%) of the Baseline value.

Table 16. Results of Feature Expansion+SMOTE with Tweet+Wikipedia Corpus

Model	Accuracy (%)			
	CNN	GRU	GRU-CNN	CNN-GRU
Baseline	**91.14**	**90.73**	**90.57**	**91.20**
Baseline+Top1	96.86 (+6.28)	95.03 (+4.71)	95.52 (+5.46)	97.52 (+6.92)
Baseline+Top5	**96.96 (+6.39)**	95.39 (+5.14)	95.23 (+5.14)	**97.58 (+6.99)**
Baseline+Top10	96.82 (+6.23)	**95.55 (+5.31)**	**95.69 (+5.65)**	97.23 (+6.61)

The test results in Tables 14, Table 15 and Table 16 show that the accuracy value against the baseline has increased significantly in all classifier models and it can be said that the feature+smote expansion test has succeeded in increasing the accuracy value significantly.

3.2. Discussion

Based on the test results in Table 8, the best accuracy value is obtained by the basic Allgram feature with the labeling by system dataset on both the CNN model and also the GRU, with a fairly large accuracy value of 91.14% and also 90.73%. based on these results, the baseline used for further testing is Allgram using dataset by system.

Furthermore, testing of feature selection is carried out to determine how many maximal TF IDF features have the highest accuracy value results. The test results in Table 9 showed that there is no significant increase in accuracy. It can be seen that for each max features tested, starting from 2000 features up to 10000, there is an increase in accuracy of (+0.86%) and (+0.16). However, in testing 15000 features, there is a decrease in accuracy. With this it can be concluded that, the more selection of max TF-IDF features does not guarantee a significant increase in accuracy.

Based on these tests, the baseline that will be used for feature expansion testing is the granularity labeling dataset on both the CNN model and also the GRU, with an accuracy value of 91.14% and also 90.73%. The test results in Table 11 showed that there is an increase in accuracy values in almost all classifier models. An increase in accuracy is found in each classifier model, CNN gets an increase in accuracy values of (+0.12%) at baseline+top5, GRU gets an increase in accuracy values of (+0.03 %) at baseline+top1, and GRU-CNN gets an increase in accuracy value of (+0.57%).

The feature expansion built with corpus similarity to the Tweet dataset tends to experience a decrease in accuracy for all classifier models as the top-n features enlarge, as shown in Table 13. The greater the top-n value to the baseline, the greater the decrease in accuracy. The results of testing the feature expansion built with the Tweet+wikipedia corpus similarity have an increase in the accuracy value, although it is not significant. It can be seen that the GRU-CNN classifier model on Baseline+Top1 has an increase in accuracy

value of (+0.01%), the CNN-GRU classifier model on Baseline+Top1 has an increase in accuracy value of (+0.19%), and on the CNN-GRU model on baseline+top10 with an increase in accuracy value of (+0.07%).

Testing the feature+smote expansion with the Wikipedia corpus, tweets, and tweet+wikipedia shows that the accuracy value against the baseline has increased significantly. The feature+smote expansion is carried out with the same corpus as the previous feature expansion test using 3 corpus namely Wikipedia corpus, tweet, and tweet+wikipedia. From the results of the feature+smote expansion testing we can see in Tables 14, Table 15 and Table 16. The highest accuracy value obtained on the Wikipedia corpus is the baseline+top1 test on the CNN-GRU model with an accuracy value of 97.77% (+7.12%) of baseline value. By comparing other studies that also use the combined CNN-GRU method [8][9], the use of fastText word embedding can provide a higher accuracy value as found in this study. Summary of all scenarios results are shown on Fig. 3.

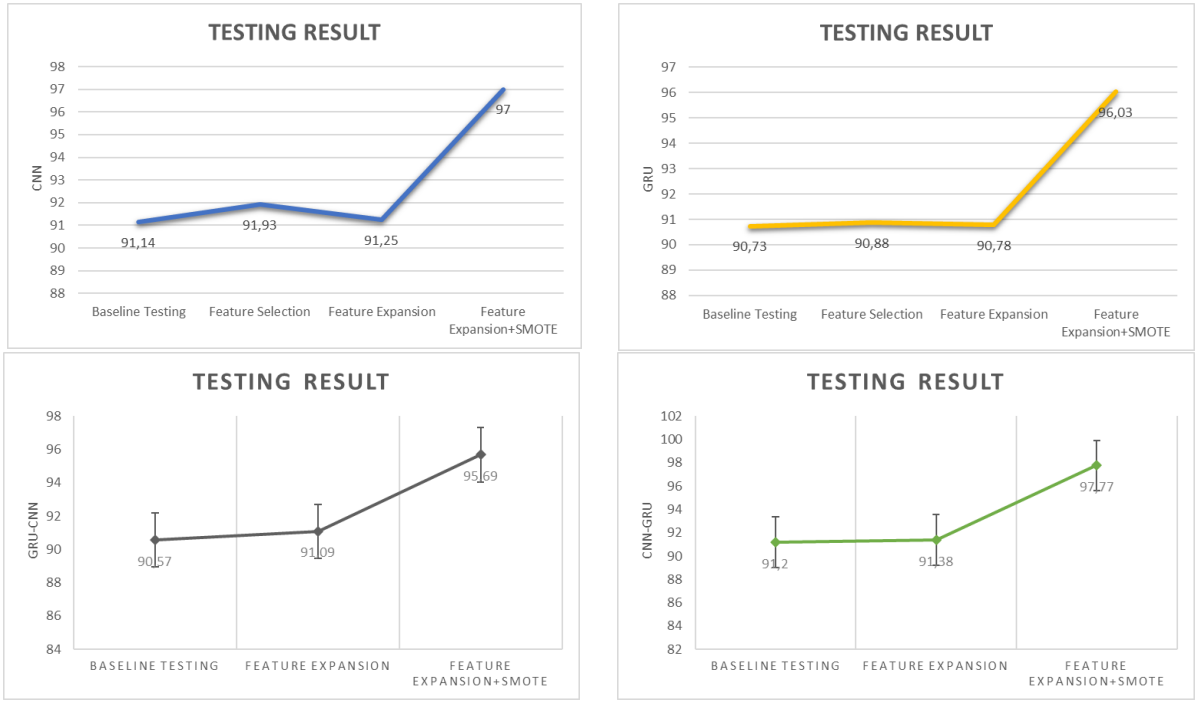

Fig. 3. Testing results of All Scenarios (%)

4. CONCLUSION

This research has conducted sentiment analysis for the Indonesian language tweet dataset, using several test scenarios. The test scenarios carried out include determining the baseline model, then carrying out feature extraction and selection on TF-IDF, then testing feature expansion against the baseline. In the first test to determine the baseline, the best results obtained on the 80:20 data split are the number of TF-IDF features of 10,000 features. The last test is the expansion of features+smote with fastText to test whether adding data balancing using SMOTE can make the accuracy value increase significantly.

The scenario that has the most significant impact is obtained on the Wikipedia corpus using the baseline+top1 scenario on the CNN-GRU model with an accuracy value of 97.77% which increases the accuracy value by 7.12% against the baseline values.

Based on these tests, it can be concluded that sentiment analysis research on Twitter social media using the Convolutional Neural Network and Gated Recurrent Unit methods can produce fairly high accuracy, and feature expansion testing of the deep learning model can provide a significant increase in accuracy values.

Suggestions for further research is to conduct a survey of sentiment words with a larger number of corpus, so that the distribution of Granularity labeling can be evenly distributed and provide a balanced dataset for the input in deep learning models that can impact models performance.

REFERENCES

[1] J. D. Kelleher, B. M. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*, MIT Press, 2020, https://books.google.co.id/books?id=UM_tDwAAQBAJ.

[2] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning : Principles and Techniques*. CRC Press, 2019, https://doi.org/10.1201/9780429469275.
[3] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps*, Apress, 2017, https://doi.org/10.1007/978-1-4842-2866-1.
[4] Y. Santur, "Sentiment analysis based on gated recurrent unit," *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*, pp. 1-5, 2019, https://doi.org/10.1109/IDAP.2019.8875985.
[5] Y. Pan and M. Liang, "Chinese Text Sentiment Analysis Based on BI-GRU and Self-attention," *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*, pp. 1983–1988, 2020, https://doi.org/10.1109/ITNEC48623.2020.9084784.
[6] M. Sultan Mahmud *et al.*, "A Hierarchical multi-input and output Bi-GRU Model for Sentiment Analysis on Customer Reviews," *IOP Conf Ser Mater Sci Eng*, vol. 322, no. 6, p. 62007, 2018, https://doi.org/10.1088/1757-899X/322/6/062007.
[7] O. B. Deho, W. A. Agangiba, F. L. Aryeh, and J. A. Ansah, "Sentiment analysis with word embedding," *IEEE International Conference on Adaptive Science and Technology, ICAST*, pp. 1-4, 2018, https://doi.org/10.1109/ICASTECH.2018.8506717.
[8] A. Zouzou and I. el Azami, "Text sentiment analysis with CNN GRU model using GloVe," *5th International Conference on Intelligent Computing in Data Sciences, ICDS 2021*, pp. 1-5, 2021, https://doi.org/10.1109/ICDS53782.2021.9626715.
[9] N. Habbat, H. Anoun, and L. Hassouni, "Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model," *IEEE Engineering Management Review*, 2022, https://doi.org/10.1109/EMR.2022.3208818.
[10] Y. Cheng, L. Yao, G. Xiang, G. Zhang, T. Tang, and L. Zhong, "Text Sentiment Orientation Analysis Based on Multi-Channel CNN and Bidirectional GRU with Attention Mechanism," *IEEE Access*, vol. 8, pp. 134964–134975, 2020, https://doi.org/10.1109/ACCESS.2020.3005823.
[11] Y. Cheng *et al.*, "Sentiment Analysis Using Multi-Head Attention Capsules with Multi-Channel CNN and Bidirectional GRU," *IEEE Access*, vol. 9, pp. 60383–60395, 2021, https://doi.org/10.1109/ACCESS.2021.3073988.
[12] W. Li, Y. Xu, and G. Wang, "Stance Detection of Microblog Text Based on Two-Channel CNN-GRU Fusion Network," *IEEE Access*, vol. 7, pp. 145944–145952, 2019, https://doi.org/10.1109/ACCESS.2019.2944136.
[13] L. xia Luo, "Network text sentiment analysis method combining LDA text representation and GRU-CNN," *Pers Ubiquitous Comput*, vol. 23, no. 3–4, pp. 405–412, 2019, https://doi.org/10.1007/s00779-018-1183-9.
[14] M. S. Başarslan and F. Kayaalp, "MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis," *Journal of Cloud Computing*, vol. 12, no. 1, 2023, https://doi.org/10.1186/s13677-022-00386-3.
[15] C. N. Dang, M. N. Moreno-García, and F. de La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, 2021, https://doi.org/10.1155/2021/9986920.
[16] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, 2020, https://doi.org/10.1109/ACCESS.2020.2969854.
[17] A. Rahman and M. S. Hossen, "Sentiment Analysis on Movie Review Data Using Machine Learning Approach," *2019 International Conference on Bangla Speech and Language Processing (ICBSLP),* pp. 1-4, 2019, https://doi.org/10.1109/ICBSLP47725.2019.201470.
[18] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion for sentiment analysis in twitter," *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 509–513, 2018, https://doi.org/10.1109/EECSI.2018.8752851.
[19] N. C. Dang, M. N. Moreno-García, and F. de la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics,* vol. 9, no. 3, p. 483, 2020, https://doi.org/10.3390/electronics9030483.
[20] A. P. Pandian, "Performance evaluation and comparison using deep learning techniques in sentiment analysis," *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 2, pp. 123-134, 2021, https://irojournals.com/jscp/V3/I2/06.pdf.
[21] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: an overview," *Science China Information Sciences*, vol. 63, no. 1, pp. 1–36, 2020, https://doi.org/10.1007/s11432-018-9941-6.
[22] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif Intell Rev*, vol. 53, no. 6, pp. 4335–4385, 2020, https://doi.org/10.1007/s10462-019-09794-5.
[23] I. K. Candradinata and E. B. Setiawan, "Analisis Sentimen Pada Twitter Mengenai Layanan Toko Online Dengan Metode Naive Bayes," *eProceedings of Engineering*, vol. 7, no. 3, 2020, https://doi.org/10.34818/eoe.v7i3.14232.
[24] I. Alkahfi and K. Chiuloto, "Penerapan Model Gated Recurrent Unit Pada Masa Pandemi Covid-19 Dalam Melakukan Prediksi Harga Emas Dengan Menggunakan Model Pengukuran Mean Square Error," *Prosiding SNASTIKOM: Seminar Nasional Teknologi Informasi & Komunikasi*, 2021, https://garuda.kemdikbud.go.id/documents/detail/2573840.
[25] F. Hamdi Bahar, N. Indah Sari, and dan Armin Lawi, "Klasifikasi Suara Kucing dan Anjing Menggunakan LSTM-GRU dan ANN-BP," *Proceeding KONIK (Konferensi Nasional Ilmu Komputer)*, vol. 5, pp. 202–207, Aug. 2021, Accessed: Feb. 04, 2023. [Online]. Available: https://prosiding.konik.id/index.php/konik/article/view/51.

[26] N. Shehab, M. Badawy, and H. Arafat, "Big data analytics and preprocessing," *Machine learning and big data analytics paradigms: analysis, applications and challenges*, pp. 25-43, 2021, https://doi.org/10.1007/978-3-030-59338-4_2.
[27] R. A. A. Malik and Y. Sibaroni, "Multi-aspect Sentiment Analysis of Tiktok Application Usage Using FasText Feature Expansion and CNN Method," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 277–285, 2022, https://doi.org/10.47065/josyc.v3i4.2033.
[28] H. R. Alhakiem, and E. B. Setiawan, "Aspect-Based Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 6, no. 5, pp. 840-846, 2022, https://doi.org/10.29207/resti.v6i5.4429.
[29] M. A. Raihan, and E. B. Setiawan, "Aspect Based Sentiment Analysis with FastText Feature Expansion and Support Vector Machine Method on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 6, no. 4, pp. 591-598, 2022, https://doi.org/10.29207/resti.v6i4.4187.
[30] S. D. Lestari and E. B. Setiawan, "Sentiment Analysis Based on Aspects Using FastText Feature Expansion and NBSVM Classification Method," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 469–477, 2022, https://doi.org/10.47065/josyc.v3i4.2202.
[31] R. A. Yahya and E. B. Setiawan, "Feature Expansion with FastText on Topic Classification Using the Gradient Boosted Decision Tree on Twitter," *2022 10th International Conference on Information and Communication Technology (ICoICT)*, pp. 322-327, 2022, https://doi.org/10.1109/ICoICT55009.2022.9914896.
[32] D. Roy, D. Ganguly, S. Bhatia, M. Mitra, "Using word embeddings for information retrieval: How collection and term normalization choices affect performance," *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1835–1838, 2018, https://doi.org/10.1145/3269206.3269277.
[33] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, "An evaluation of SVM and naive bayes with SMOTE on sentiment analysis data set," *ICEAST 2018 - 4th International Conference on Engineering, Applied Sciences and Technology: Exploring Innovative Solutions for Smart Society*, pp. 1-5, 2018, https://doi.org/10.1109/ICEAST.2018.8434401.
[34] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognit Lett*, vol. 150, pp. 250–257, 2021, https://doi.org/10.1016/j.patrec.2021.07.009.
[35] W. Satriaji and R. Kusumaningrum, "Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis," *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, pp. 99–103, 2018, https://doi.org/10.1109/ICICOS.2018.8621648.
[36] E. Loginova, W. K. Tsang, G. van Heijningen, L. P. Kerkhove, and D. F. Benoit, "Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data," *Machine Learning*, pp. 1-24, 2021, https://doi.org/10.1007/s10994-021-06095-3.
[37] S. Kedas, A. Kumar, and P. K. Jain, "Dealing with Class Imbalance in Sentiment Analysis Using Deep Learning and SMOTE," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 106, pp. 407–416, 2022, https://doi.org/10.1007/978-981-16-8403-6_37.
[38] Hermanto, A. Y. Kuntoro, T. Asra, E. B. Pratama, L. Effendi, and R. Ocanitra, "Gojek and Grab User Sentiment Analysis on Google Play Using Naive Bayes Algorithm and Support Vector Machine Based Smote Technique," in *Journal of Physics: Conference Series*, vol. 1641, no. 1. 2020, https://doi.org/10.1088/1742-6596/1641/1/012102.
[39] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, https://doi.org/10.1109/ACCESS.2022.3149482.
[40] A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. ul Haq, "Understanding of Convolutional Neural Network (CNN): A Review," *International Journal of Robotics and Control Systems*, vol. 2, no. 4, pp. 739–748, 2022, https://doi.org/10.31763/ijrcs.v2i4.888.
[41] N. Aslam, F. Rustam, E. Lee, P. B. Washington, and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," *IEEE Access*, vol. 10, pp. 39313–39324, 2022, https://doi.org/10.1109/ACCESS.2022.3165621.
[42] A. R. Khan, A. T. Khan, M. Salik, and S. Bakhsh, "An Optimally ConFig.d HP-GRU Model Using Hyperband for the Control of Wall Following Robot," *International Journal of Robotics and Control Systems*, vol. 1, no. 1, pp. 66–74, 2021, https://doi.org/10.31763/ijrcs.v1i1.281.

BIOGRAPHY OF AUTHORS

Ahmad Zahri Ruhban Adam, is currently pursuing a bachelor's degree in computer science at Telkom University, Indonesia. Email: zahriadam@student.telkomuniversity.ac.id

Erwin Budi Setiawan, is a senior lecturer in School of Computing, Telkom University, Bandung, Indonesia. He has more than 10 years Research and Teaching experience in the domain of Informatics. Currently, he is a Associate Professor. His research interests are machine learning, people analytic, and social media analysis. Email: erwinbudisetiawan@telkomuniversity.ac.id