Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

Topic Detection on Twitter using GloVe with Convolutional Neural Network and Gated Recurrent Unit

Moh Adi Ikfini M, Erwin Budi Setiawan*

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]adiikfini@student.telkomuniversity.ac.id, [2,*]erwinbudisetiawan@telkomuniversity.ac.id
Correspondence Author Email: erwinbudisetiawan@telkomuniversity.ac.id
Submitted: **09/08/2023**; Accepted: **25/09/2023**; Published**: 27/09/2023**

Abstract−Twitter is a social media platform that allows users to share thoughts or information with others for all to see. However, twitters often use abbreviations, slang, and incorrect grammar because tweets are limited to 280 characters. Topic detection often has problems with low accuracy, one method that can be used to overcome this problem is feature expansion. Feature expansion on Twitter is a semantic addition to the process of expanding the original text syllables to make it look like a large Document. That way, feature expansion is used to reduce word mismatches. This study uses the expansion of the GloVe feature with the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) classification methods. The results show that the topic detection system with the GloVe feature extension and CNN-GRU hybrid classification has an accuracy of 94.41%.

Keywords: Twitter; Feature Expansion; GloVe; CNN; GRU

1. INTRODUCTION

Twitter is a social media platform that enables its users to send and read short text-based messages, limited to a maximum of 280 characters tweet [1]. Over time, Twitter has evolved into one of the largest sources of information, providing a convenient, rapid, and reliable platform for users to share anything happening around them with friends and followers Consequently, Twitter can be utilized for conducting research on topic detection. Topic detection has been previously explored using techniques such as Word Embedding [2].

Topic detection is the process of assigning subjects to unstructured data or text. Classifying topics for documents (text), it allows for searches, statistical analysis, and meaningful classification [3]. In this research, topic detection is employed to retrieve a series of text documents (tweet corpus) and provide a set of topics for analysis, describing the content of each tweet in the corpus. There are challenges in categorizing tweets based on topics due to the character limit of 280 characters, leading to issues with grammar, slang, and low-quality tweets [2]. The utilization of slang terms in tweets can cause variations in vocabulary, making it difficult to comprehend the intended meaning [4]. To address this problem, one approach is using feature expansion. Feature expansion is a technique that enriches the original text by adding a semantic component to obtain a broader perspective of the text document [5].

Related research on topic detection using expansion features, Yahya et al [3], conducted research on detecting topics on Twitter using the Gradient Boosted Decision Tree method. In this study, the authors used the extraction of the TF-IDF (Term Frequency Inverse Document Frequency) feature and the expansion of the fastText feature, with an accuracy of 91.39% and an F1 score of 91,44%. Twitter used data of 30360 data and a news corpus of 97.794 data.

In addition, Ramadhy et al [4], did some research to expand the feature set. Using FastText and Logistic Regression, this study recognizes Trending Topic Analysis on Twitter social media. This study has an accuracy rate of 76.39%. However, compared to the baseline, the accuracy of this study increased by 14%. From a basic accuracy of 76.25%, after implementing feature expansion.

In the case of hybrid methods, Xu B et al [5], High Performance Web Attack Detection Method based on the CNN-GRU Model, a hybrid experiment between CNN and GRU using word2vec feature extraction, This hybrid layer is created to increase the accuracy of a single method, with CNN as the input layer and GRU as the output layer, the accuracy results reach 97, 56%, compared to CNN only 93.44% and GRU has a yield of 97.17%.

Cao B et al [6], it was shown that the CNN-GRU model with the CSIC 2010 dataset conducted an experiment on the Network Intrusion Detection Model Based on CNN and GRU showing increased accuracy compared to the single model. Using the NSL-KDD dataset, the CNN-GRU hybrid method has an accuracy of 99.69%, superior to CNN and GRU as a single model.

Feature expansion for sentiment analysis in Twitter [7]. in this study various tests were carried out on feature expansion with satisfactory results to increase the level of accuracy as well as the function of feature expansion to expand text documents which results in the system being more accurate in checking data, in this study This was tested using TF-IDF as a feature expansion method with logistic regression [8], NB, and SVM as the model, and the accuracy results reached 82.02%.

Based on the explanation previously described, so far no research has been found that has developed a Topic Detection system using the extended GloVe feature with the CNN-GRU hybrid method using datasets obtained from Twitter . The tweet data obtained generally contains comments in text form. Then, the tweet data will go through a feature expansion process to make it more accurate in categorizing tweets on topics based on the words loaded in the tweet.

The purpose of this research is to build and analyze a topic detection system using GloVe feature expansion with CNN-GRU which is tested by comparing single models, hybrid models, single-model feature expansion, and

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

hybrid-feature expansion to find out how much the accuracy improvement is using a dataset in the form of text obtained from Indonesian Twitter.

The organizational structure of this research paper is as follows: section 2 describes the methods used in this research, section 3 describes the results and discussion, and section 4 contains conclusions.

2. RESEARCH METHODOLOGY

2.1 System Design

The model design in this study can be seen in **Fig 1**.

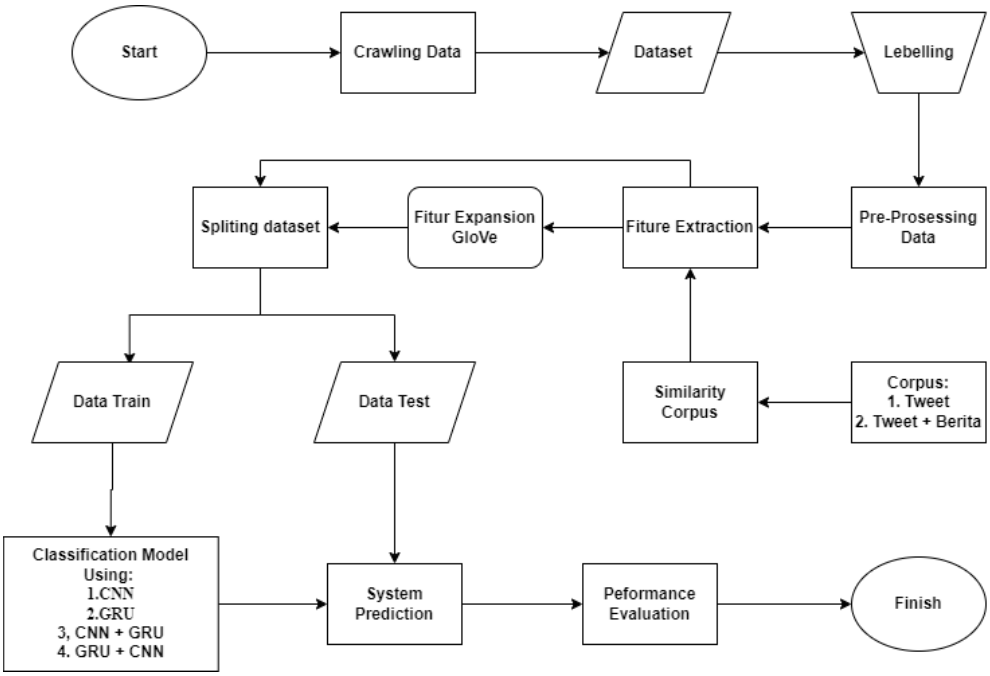

Figure 1. Topic Detection System

This section delineates the approach employed in this study. As depicted in **Figure 1**, the system operates through the sequence outlined in the subsequent flowchart. The methodology in this paper comprises 5 phases, namely data preprocessing, feature extraction, feature expansion, classification model, and evaluation.

2.2 Data Crawling

Data was obtained using the Python programming language snscrape library [9]. This research relies on data from Twitter in Indonesian. Twitter data is collected using certain keywords for tweet searches. The crawl procedure will collect data for 1000 tweets at a time. The data used was collected between January 2022 and March 2023. A total of 56,236 Twitter datasets were acquired. However, due to duplication of data, the number of datasets is reduced, and the final dataset before the preprocessing process is 55,411. The distribution of crawling data is shown in **Table 1**.

Table 1. Data Distribution

Label	Amount	Percentage
Business	7.190	13%
Health	6.030	11%
Sport	5.978	11%
Education	5.898	10%
Automotive	5.718	10%
Entertainment	5.527	10%
Economy	5.384	10%
Lifestyle	4.905	9%
Technology	4.803	8%
Travel	3.978	8%
Total	**55.411**	**100%**

2.3 Labelling

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

Data labeling is done on the data set before the classification process. This tagging is done to distinguish the topic of each tweet in the dataset. In this research, there are ten different labels for each tweet. These categories are based on the most popular themes that frequently occur on many Indonesian News portal. Each topic that appears often on multiple news websites is used as a reference label to categorise topics in Twitter data. Following an assessment of the most popular themes on Indonesian news portals, ten labels were chosen for use in this study. The label used Business, Health, Sports, Education, Automotive, Entertainment, Economy, Lifestyle, Technology, and Travel. In this study, the process of manually labelling data [10]. To ensure the accuracy of the labelling results, each data is checked by at least three people. The method of majority votes is used in decision-making when there is disagreement in the marking process [11]. Examples of labels can be seen in **Table 2**.

Table 2. Example of a label on a tweet

Tweet	Label
Perbedaan antara kedua definisi tersebit adalah yaitu perencanaan mempersingkat waktu layanan berdasarkan variabel yang dapat dikendalikan dan antara pelayanan yg menurunkan persedian,waktu tunggu pelanggan dan tingkat produksi hingga kendali bisnis kepada manajemen.	Business
@irmaalazmi klu mau disesuaikan ama budget mendingan ikut travel emang ngk bebas tp ngk semahal private traveling sama perbanyak amal ibadah buat menang raffle klu kudu beli calo ya nabungnya double ðŸ˜€	Travel
Bukan cuma eActros LongHaul, Mercedes Benz juga masih menguji truk jarak jauh dengan basis Fuel Cell Electric Vehicle (FCEV) dengan Hidrogen https://t.co/SRhtf2llUm	Automotive

2.4 Data Preprocessing

Data preprocessing is the preparation of data before it is processed. Raw, incomplete, or inconsistent data will be translated into a machine-readable format [12]. Text preprocessing is the processing and preparation of text data before analysis. These steps include data cleaning, converting letters into lowercase, dividing the text into words or tokens, filtering or removing irrelevant or meaningless words, and changing words into their basic form by deleting affixations [13], [14]. To help in the preprocessing process, the NLTK (Natural Language Toolkit) Python library enables tagging, stemming, classification, tokenization, and Capabilities for parsing and semantic analysis [15], in addition to using the Pysastrawi library to assist in the stemming process [14].

Data cleaning is used to clean text from unwanted characters or emotes because tweet data is not entirely in text form. Some tweets use hashtags, mentions, website URLs, and other unique characters [15]. As a result, text that meets these requirements is cleaned up first. Mentions, hashtags, HTTP links, special characters, double spaces, and movie titles are some procedures [16], [17].

Then the next thing is stemming. Stemming is the process of separating affixes, namely prefixes, infixes, suffixes, and confixes (combined prefixes and suffixes), where words are derived to become essential. With stemming, word variations with the same root word will be considered as the same token (feature) [18]. followed by tokenization is the process of dividing sentences into words, phrases, and symbols called tokens [18]. The generated tokens will help in data parsing and processing. Developments in tokenization in this context separate character series into basic processing units and interpret and classify isolated tokens to form higher level tokens. Next, the raw text is processed and divided into smaller units [19].

2.5 Feature Extraction

Feature extraction is converting (or extracting) text documents into features that machine learning classification systems can process fast. One of the most essential approaches in data mining and text classification is feature extraction, which generates feature values in documents [20].

This study uses TF-IDF (Term Frequency-Inverse Document Frequency) as a feature extraction method. TF-IDF computes the weight or value of each word (token) in a corpus document. This method is frequently utilized in information retrieval and text creation to evaluate the relationship of each word in a relationship document [21]. This normalization process determines the weight of terms that appear frequently in a document. The document is translated into a weight based on the number of occurrences [22]. This approach determines the frequency with which this word appears in the Twitter document. The formula for TF-IDF is as follows:

$$w_{ij} = tf_{ij} \times IDF_j, with\ IDF_j \left(\log \frac{N}{df}\right) \quad (1)$$

The TF-IDF formula is used to calculate the weight of words in a document. The TF-IDF weight W_{ij} is obtained by multiplying the Term Frequency tf_{if} with the Inverse Document Frequency IDF_j of that document. TF tf_{ij} represents how often word i appears in document j, while IDF IDF_j measures how rarely word i appears across the entire document collection (corpus) based on log(N/df), where N is the total number of documents in the collection, and df is the number of documents that contain word i. Thus, TF-IDF assigns a higher weight to words that frequently appear in the specific document but rarely appear in the entire document collection, indicating their importance in the context of that document.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

2.6 Feature Expansion

GloVe is a log-bilinear model or can be called a calculation-based model. GloVe studies word relationships by counting how often words appear with each other in a given corpus. The probability occurrence ratio of words has the potential to encode several forms of meaning and help improve performance on word analogy problems [23].
The goal of the GloVe method word representation process is to obtain semantic links between words based on a matrix of common occurrences [23]. GloVe employs a global matrix factoring method, which involves constructing a matrix that encodes the presence or absence of words in a document [24]. GloVe investigates word relationships by calculating how frequently the words appear together in a given corpus. The likelihood ratio of word appearance can convey meaning and aid in word analogy problem performance [24]. The corpus of this study consists of tweets and IndoNews. The tweet corpus uses tweet data, while the tweet + IndoNews corpus is constructed using the combined tweet+IndoNews. **Table 3** is an example of the similarity of the word "hujan" based on the corpus of Tweets with the top 15 similarities.

Table 3. Example of Top 15 Similarity of "hujan" words based on Corpus Tweets + IndoNews

hujan	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
	deras	intensitas	ringan	guyur	meteor
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	gerimis	lebat	curah	awan	sang
	Rank 11	**Rank 12**	**Rank 13**	**Rank 14**	**Rank 15**
	sekolah	rintik	Jabodetabek	cuaca	wekeri

2.7 Classification Model

2.7.1 CNN

Convolutional Neural Networks (CNN) are robust natural language processing architectures, particularly for topic detection in text. CNN can find local patterns and extract significant information from text using the convolution layer. Because of hierarchical implementation, CNN can grasp the context and links between words in documents, making it particularly good at finding topics[6]. CNN's ability to automatically extract features and adapt to significant texts further improves its accuracy and adaptability. As a result, CNN has been widely employed in tasks such as text classification, sentiment analysis, and news categorization, considerably contributing to text analysis and online content understanding [5]. The CNN architecture comprises main layers; feature extraction (convolutional and pooling) and fully connected layers. The convolutional layer processes small input parts through filters, performing multidimensional matrix calculations. The output goes through the pooling layer, reducing samples and carrying information to the next layer. Finally, the feature extraction layer produces activation maps, reshaped into vectors for the fully connected layer input.

Table 4. Parameters setting for CNN

Parameters of Model	
Number of convolution	64
Size of kernel	3
Stride of convolution operation	2
Padding	SAME
Activation	RULE
Size of local max pooling	2
Dropout	0.1
Dance, Activation	32, RULE
Flatten	-
Dance, Activation	10, SOFTMAX

The parameters needed to implement CNN are shown in Table 4. CNN has a convolutional layer with 64 filters with a width of 3 and a local max-pooling layer of size 2, using the SAME padding technique and the RELU activation function. GRU has 64 units with a drop rate of 0.1. The Flatten layer converts the output data from convolution and unification into 1D form before being forwarded to the Dense layer. And ends Dance with unit 10, softmax activation as the output.

2.7.2 GRU

Gated Recurrent Unit (GRU) is a recurrent neural network (RNN) type used in text analysis and in-text topic recognition. GRU solves the vanishing gradient problem in RNNs by regulating information flow with a gate mechanism. This allows GRU to detect effective long-term patterns in text and is a popular choice for natural language processing applications like text categorization and sentiment analysis [25]. GRU can sequentially process data, making it an effective tool for addressing complex texts with sequential structures [26].

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

The GRU has unique gate mechanisms, including input gates, update gates, and reset gates, which regulate the flow of information within the cell and allow for precise control of relevant and irrelevant information. This mechanism helps overcome the vanishing gradient problem that often occurs in traditional RNNs, making GRU more efficient at remembering remote contexts in long sequence data. GRU has been used extensively in various applications, including language modelling, machine translation, and time series analysis, due to its ability to deal with sequence problems.

Table 5. Parameters setting for GRU

Parameters of Model	
Number of GRU Unit	32
Return_sequences	TRUE
Kernel_regularizer	L2(0.1)
Dropout	0.6
Dance, Activation	32, RULE
Flatten	-
Dance, Activation	10, SOFTMAX

Table 5, shows the Gated Recurrent Unit (GRU) layer. First, the GRU layer with the number of hidden units 32 is used to process sequence data by managing the flow of information through the gate mechanism, including the input gate, update gate, and reset gate. The GRU layer generates a sequence output for each timestep with the parameter return_sequences=True. A dropout is applied to the GRU layer during training to prevent overfitting with a probability of 0.6. Then, the output from the GRU layer is connected to the Dense layer with 32 units and the ReLU activation function. After that, Flatten is performed to convert the data into 1D vector form before being passed to the last Dense layer with 10 units and the softmax activation function to generate class probabilities in multi-class classification tasks with 10 different classes.

2.7.3 Hybrid

A hybrid model is a system or model development methodology that combines two or more separate methodologies or strategies to meet specific goals. A hybrid model combines many methods or methodologies used to improve system performance, accuracy, or efficiency in machine learning and data analysis [27]. The primary goal of employing a hybrid model is to use each method's benefits while overcoming each method's shortcomings. As a result, the hybrid model can deliver better and more optimal performance in handling complicated issues that are difficult to tackle with a single method alone [28].

Table 6. Parameters setting for CNN-GRU

Parameters of Model	
Number of convolution	32
Size of kernel	3
Stride of convolution operation	2
Padding	SAME
Activation	RULE
Kernel_regularizer	L2(0.1)
Dance, Activation	32, RULE
Size of local max pooling	2
Number of GRU Unit	32
Return_sequences	TRUE
Flatten	-
Dance, Activation	10, SOFTMAX

Table 6, show a combination of the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) layers. The model uses a Conv1D layer with 32 filters, a kernel size of 3, and the ReLU activation function, and implements L2 regularization with a power of 0.01. After that, proceed with a Dense layer of 32 units and activation of ReLU. MaxPooling1D is done with 2 pool sizes and 2 steps to reduce the spatial dimension of the data. Furthermore, the GRU layer with the number of hidden units 32 processes sequence data with return_sequences=True to produce output sequences for each timestep. The output data from GRU is then averaged into a 1D vector before passing through the last Dense layer with 10 units and softmax activation to perform multi-class classification with 10 different classes. This model can well extract features from sequence data using convolution layers and also consider sequence dependencies with GRU layers for more complex classification tasks.

Table 7. Parameters setting for GRU-CNN

Parameters of Model	
Number of GRU Unit	32

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

Return_sequences	TRUE
Number of convolution	32
Size of kernel	3
Stride of convolution operation	2
Padding	SAME
Activation	RULE
Kernel_regularizer	L2(0.1)
Dance, Activation	32, RULE
Size of local max pooling	2
Flatten	-
Dance, Activation	10, SOFTMAX

Table 7, show a combination of the Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN) layers. The model starts with the GRU layer with hidden unit number 32 which will process sequence data with parameter return_sequences=True, producing sequence output for each timestep. This is followed by a Conv1D layer with 32 filters, a kernel size of 3, and the ReLU activation function, and L2 regularization is applied to reduce overfitting. After the convolution layer, it is continued with a Dense layer of 32 units and the ReLU activation function. MaxPooling1D is done with pool size 1 and step 2 to reduce the spatial dimension of the data. Next, the data is flattened into a 1D vector before being passed to the final Dense layer with 10 units and the softmax activation function to perform multi-class classification with 10 different classes. This model combines the advantages of GRU in processing sequence data and CNN's ability to extract spatial features from data, making it suitable for classification tasks on sequence data with relevant spatial structures.

With the advantages of each topic detection is a compelling reason to use CNN-GRU and GRU-CNN as an architecture, because Convolutional Neural Networks (CNN) effectively extract local features from text and identify relevant patterns. At the same time, Gated Recurrent Units (GRU) can understand the context of long-term text sequences.

2.8 Evaluasi

The value of accuracy is used in this study to calculate the system's performance. Accuracy is calculated using four terms from the confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). However, because this is a multiclass problem, accuracy indicates the percentage of correctly classified data as a percentage of total data [29]. The equation determines the accuracy value:

$$Accuracy(y, \overline{y}) = \frac{1}{n_{sample}} \sum_{i=0}^{n_{sample}-1} 1(y, \overline{y}) \quad (2)$$

3. RESULT AND DISCUSSION

This research outlines four different scenarios that serve as the basis for topic classification. This approach involves utilizing Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) for classification models, TF-IDF N-grams for initial testing and feature extraction, and GloVe embedding for feature enrichment. The scenario is constructed as follows: In the first scenario, the split ratio plays an important role. This involves partitioning the data in some way to assess the impact on the classification results. The effectiveness of this approach determines the experiment for the next scenario. Moving on to the second scenario, baseline comparisons are performed, comparing model performance against established standards. The results from the first scenario guide the selection of the best performing method for this step.

Taking the best results from the second scenario, the third scenario takes advantage of enhanced GloVe functionality. By integrating advanced GloVe techniques, the aim is to refine and optimize the basic results achieved so far. Finally, the fourth scenario uses a hybrid approach. A new classification model is introduced, combining elements of the CNN and GRU architectures. This hybrid model combines with an expanded GloVe feature set to further enhance the best baseline results. In summary, this study employs systematic scenario development, leveraging multiple techniques including CNN, GRU, TF-IDF N-gram, and GloVe embedding, to comprehensively explore and improve thematic classification results. Each scenario builds on the insights gained from the previous one, leading to a progressively refined and nuanced classification framework.

3.1 Result

3.1.1 Scenario 1

The first test is carried out to determine the baseline by selecting the appropriate data split scenario and the number of features to be used in TFIDF. The split ratio used in this study is 90:10, 80:20, and 70:30. The test results can be seen in **Table 8** below.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

Table 8. Result of First Scenario

Split Ratio	Accuracy (%)	
	CNN	GRU
90:10	92.11	92.76
80:20	**92.15**	**93.46**
70:30	91.82	93.30

Table 8 shows the best accuracy result is the splitting ratio of 80:20 with the accurate value on the CNN model of 92.15% and the GRU model of 93.46%. The best results from the first scenario continued to the second scenario to compare the best baseline.

3.1.2 Scenario 2

The baseline comparison in this situation employs unigram, bigram, trigram, unigram + bigram, and unigram + bigram + trigram. The test results are shown in **Table 9**.

Table 9. Result of Second Scenario

Baseline	Accuracy (%)	
	CNN	GRU
Unigram (Baseline)	92.15	93.49
Bigram	71.59(-22.31)	71.83(-23.17)
Trigram	36.72(-60.15)	38.69(-58.62)
Unigram + Bigram	91.64(-0.55)	77.65(-16.94)
Unigram + Bigram+ Trigram	**92.19(+0.04)**	**93.59(+0.10)**

The table above shows that the baseline with the best accuracy results is Baseline + Bigram + Trigram, with the accurate development of the CNN model of 92,19% with an increase of 0.04% and GRU of 93,59% with an increase of 0.010%. Baseline results will continue in the third scenario.

3.1.3 Scenario 3

In the third case. Started feature expansion During this procedure, the baseline and classification model will be enhanced using features from the similarity corpus, such as Top 1, Top 5, Top 10, and Top 15. The corpus is of two types: corpus tweets and corpus Tweet + IndoNews. The CNN and GRU classification models will use both corpora. **Table 10** displays the results of the CNN model trials.

Table 10. Result of Third Scenario Using CNN

Feature	Accuracy (%)		
	Baseline	Twitter	Twitter + IndoNews
Top 1	92.15	92.31(+0.17)	92.23(+0.08)
Top 5		**92.32(+0.18)**	**92.38(+0.24)**
Top 10		92.09(-0.06)	92.19(+0.04)

Table 6 Here shows that the Top 5 has greater accuracy, having made three guesses: Top 1, Top 5, and 10. The experiment will stop when it finds the highest accuracy or the accuracy in the next trial drops. with the highest accuracy results on CNN using Corpus Twitter at 92.32% with an increase of 0.18% and Corpus Twitter + IndoNews at 92.38% with an increase of 0.24%.

Table 11. Result of Third Scenario Using GRU

Feature	Accuracy (%)		
	Baseline	Twitter	Twitter + IndoNews
Top 1	93.46	**94.18(+0.76)**	**94.20(+0.79)**
Top 5		94.08(+0.66)	94.16(+0.74)

Table 11. Here shows that the Top 1 has greater accuracy, having made three guesses: Top 1, Top 5, and 10. The experiment will stop when it finds the highest accuracy or the accuracy in the next trial drops. with the highest accuracy results on GRU using Corpus Twitter at 94.18% with an increase of 0.76% and Corpus Twitter + IndoNews at 94.20% with an increase of 0.79%.

In **Table 10** and **Table 11**, an experiment was carried out using the Glove feature expansion on the single CNN and GRU models using Corpus Twitter + IndoNews and Corpus Twitter, which was expanded on tweet data. The experiment was carried out continuously until high accuracy was found or accuracy decreased in subsequent experiments. With the highest accuracy results on CNN and GRU falling on the Glove expansion using Corpus Twitter + IndoNews, this can be covered with lots of vocabulary in the Corpus.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

3.1.4 Scenario 4

Table 12. Result of the Forth Scenario Using CNN+GRU

Feature	Accuracy (%)		
	Baseline	Twitter	Twitter + IndoNews
Top 1	92.15	**94.30(+2.33)**	94.40(+2.44)
Top 5		94.17(+2.19)	**94.41(+2.45)**
Top 10		-	94.15(+2.17)

The accuracy results of the CNN+GRU hybrid classification model using expansion and extraction features are shown in **Table 12**. The accuracy scores on the corpus Tweet of 94,30% with an increase of 2.33% and the corpus Twitter + IndoNews of 94,41% with an increase of 2.45% increased in the Top 1 and Top 5 features. We can see that the Tweet+IndoNews corpus has greater accuracy values. The accuracy of the Top 10 is decreasing.

Table 13. Result of the Forth Scenario Using GRU+CNN

Feature	Accuracy (%)		
	Baseline	Twitter	Twitter + IndoNews
Top 1	93.46	**94.25(+0.84)**	**94.34(+0.94)**
Top 5		94.10(+0.68)	94.15(+0.73)

The accuracy results of the CNN+GRU hybrid classification model using expansion and extraction features are shown in **Table 13**. The accuracy scores on the corpus Tweet of 94,25% with an increase of 0.84% and the corpus Twitter + IndoNews of 94,34% with an increase of 0.94% increased in the Top 1 feature. We can see that the Tweet+IndoNews corpus has greater accuracy values. The accuracy of the Top 10 is decreasing. In this scenario, all existing features and models have been implemented, namely, feature expansion which is combined with CNN-GRU and GRU-CNN hybrids, with a fairly good increase in accuracy caused by a large number of vocabularies and by their respective advantages in topic detection, the use of the CNN-GRU architecture becomes a strong choice. Convolutional Neural Networks (CNN) effectively extract local features from text and identify relevant patterns. Meanwhile, Gated Recurrent Units (GRU) can understand the context of text sequences in the long run.

3.2 Discussion

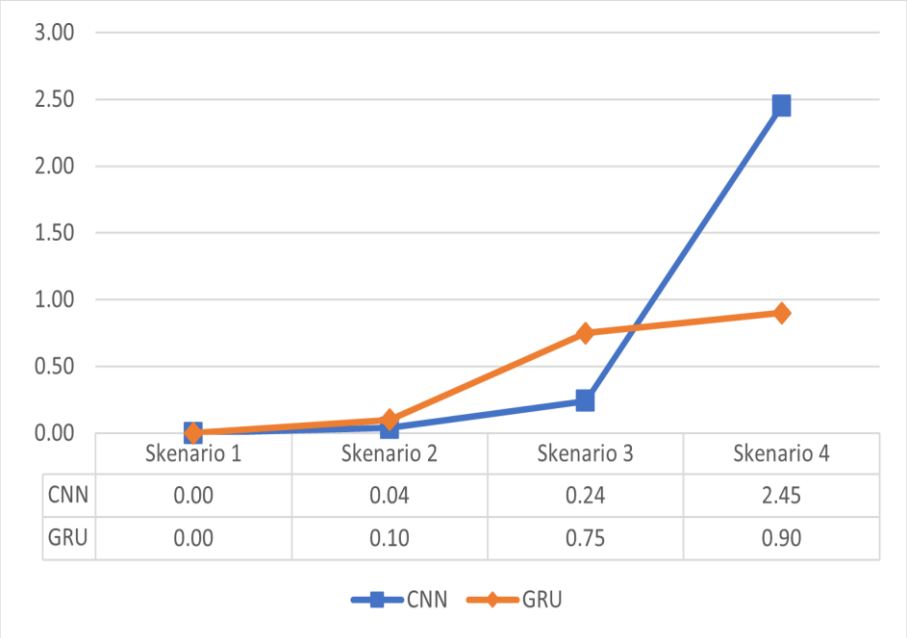

Figure 2. The Highest Relative Increase In All Scenarios

The graph results show an increase in performance in testing using the CNN, GRU, and hybrid methods with feature expansion using TF-IDF and GloVe feature extraction showed **Figure 2**. In Scenario 1, the N-gram unigram achieves the highest accuracy with a separation ratio of 80:20. In Scenario 2, the 5 N-gram test shows the highest accuracy in Unigram + Bigram and Trigram. Scenario 3 combines TF-IDF and GloVe, with GRU achieving the highest accuracy of 92.38% for Top 5 features and CNN of 94.20% for Top 5 features. In Scenario 4, the CNN+GRU hybrid outperforms the single CNN model with an increase of 2.45%, achieving an accuracy of 94.41%. Likewise, the GRU+CNN hybrid model outperforms the single GRU with an increase of 0.90%, achieving an accuracy of 94.34%.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

The improvement progress that can be witnessed in each scenario is related to many influencing factors. Starting with scenario one as the basic basis for subsequent comparisons, the observed increase, especially in scenario two, comes from variations in the use of N-grams in each iteration of the experiment. The N-gram, which denotes a sequence of N consecutive words or characters present in text or other data, offers a repertoire of advantages. These include enhanced contextual understanding, acuity of idiomatic expressions and phrases, predictive ability to anticipate next words, rudimentary sentiment analysis capabilities, identification of malicious text content, as well as efficient text compression techniques.

However, it is important to acknowledge the inherent limitations associated with N-grams. These include the intricacies of choosing the optimal size N, the constraints related to the articulation of contextual subtleties, the challenges presented by words not covered in the training corpus, and the complexities of handling rare or rare data examples. In particular, the text processing landscape has evolved rapidly with the emergence of Transformer-based models. This model has emerged as a sophisticated alternative, effectively overcoming the limitations imposed by the traditional N-gram methodology. Their distinctive strength lies in their ability to assimilate and understand a wider range of contextual cues and complex data structures. This capacity makes the Transformer models an invaluable tool for navigating the multifaceted challenges inherent in modern text analysis endeavors, demonstrating their capacity to facilitate nuanced understanding across a broad spectrum of text-based tasks.

Scenario three shows performance gains attributed to feature expansion, using the GloVe embed integration. The observed improvement in accuracy is mainly underpinned by GloVe inherent strengths, which stem from its ability to offer a holistic representation of words through combining global and local contextual information. This unique feature empowers the GloVe with the capacity to effectively deal with previously unseen or unknown words, thereby increasing its usability. In addition, GloVe facilitates the creation of word vectors that not only encapsulate semantic relations but also provide semantic operations within a mathematical framework.

The method's accessibility and computational efficiency underscore its suitability for a wide range of natural language processing tasks. Nevertheless, it is very important to acknowledge the inherent constraints of GloVe. This includes limitations in capturing highly complex contextual nuances and a potential inability to decipher deep or superior layers of meaning in highly specialized tasks. Despite these limitations, GloVe contribution remains valuable and can be widely applied in improving the representation and understanding of text data in various contexts.

The significant accuracy gain observed in scenario four is associated with the use of the CNN-GRU Hybrid Deep Learning method. This aggregation leverages its inherent advantages in understanding multimodal image and text content, resulting in complex, yet powerful feature representations for both modalities. Coupled with its proficiency in sequential data processing, the model has proven well suited for tasks requiring comprehensive analysis of image and text data, even at the expense of high model complexity and increased data requirements.

However, the CNN-GRU Hybrid Deep Learning model carries some inherent limitations. They embodies high model complexity, which can lead to complications in both training and parameter setting, in addition to the risk of overfitting due to their sophisticated architecture. The effectiveness of the model depends on the availability of large and diverse data sets, potentially limiting its applicability in scenarios with limited data resources. In addition, the complex process of selecting optimal parameters presents additional challenges.

The advantages of the Hybrid model in multimodal tasks may be contrasted with the potential for reduced performance in unimodal tasks, reflecting a trade-off between specificity and flexibility. In addition, the complexity of interpreting the decision models and underlying processes poses interpretability challenges. Careful consideration of these drawbacks vis-à-vis the specific objectives and requirements of the task at hand is essential, especially in the context of tasks involving multimodal data analysis, where the advantages of these methods can be exploited to their full potential.

4. CONCLUSION

This study used the Convolutional Neural Network (CNN) and Gate Recurrent Unit (GRU) methods to recognize themes on Twitter, with the expansion of TF-IDF N-gram extraction features as baseline and GloVe expansion features. The team analyzed 55,411 tweets. Because a tweet is a brief text message with only 280 characters, the Glove feature expansion is chosen to reduce word non-conformity caused by text data comprising multiple variations of terms, such as acronyms or slang languages. To differentiate subjects from each tweet in the dataset, ten labels are employed. showing the accuracy of the results by trying various experiments and comparing it to the baseline, and many factors influence the increase in accuracy, such as model parameters or how much it is trained. Machine learning speed, the slower the learning model, the better the performance of the model, and vice versa. However, this does not mean that the model will be studied long and produce good results. This test was carried out five times for each 1 test. As a result, the feature expansion method collaborated with hybrid is quite effective in dealing with accuracy problems in topic detection. The CNN+GRU hybrid model with TF-IDF feature extraction and GloVe feature expansion obtains an accuracy of 94.41%, a significant increase of 2.45% compared to the baseline CNN. With the same features, the GRU+CNN hybrid model obtains an accuracy of 94.34%, an increase of 0.90% from the GRU baseline. These results show that the hybrid deep learning strategy, along with feature extraction and extension, greatly improves system

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

performance while providing the highest accuracy. Proposals for further studies include experimenting with classification models, extending different features and using a larger corpus for better results.

REFERENCES

[1] P. Studi Komunikasi dan Penyiaran Islam and S. Tinggi Agama Islam As-Sunnah Deli Serdang, "Dampak Perkembangan Teknologi Informasi dan Komunikasi Terhadap Budaya Impact of Information Technology Development and Communication on Culture Daryanto Setiawan," SIMBOLIKA, vol. 4, no. 1, 2018, doi: 10.31289/simbollika.v4i1.1474.

[2] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature Expansion using Word Embedding for Tweet Topic Classification," in 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA), Denpasar: IEEE, 2016, pp. 1–5. doi: 10.1109/TSSA.2016.7871085.

[3] R. A. Yahya and E. B. Setiawan, "Feature Expansion with FastText on Topic Classification Using the Gradient Boosted Decision Tree on Twitter," in 10th International Conference on Information and Communication Technology (ICoICT), Bandung: IEEE, 2022, pp. 322–327. doi: 10.1109/ICoICT55009.2022.9914896.

[4] I. F. Ramadhy and Y. Sibaroni, "Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression," JURIKOM (Jurnal Riset Komputer), vol. 9, no. 1, p. 1, Feb. 2022, doi: 10.30865/jurikom.v9i1.3791.

[5] B. Xu and K. Mou, "A High-performance Web Attack Detection Method based on CNN-GRU Model," in 2020 IEEE 4th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2020), Chongqing, China: IEEE, 2020, pp. 804–808. doi: 10.1109/ITNEC48623.2020.9085028.

[6] B. Cao, C. Li, Y. Song, Y. Qin, and C. Chen, "Network Intrusion Detection Model Based on CNN and GRU," Applied Sciences (Switzerland), vol. 12, no. 9, May 2022, doi: 10.3390/app12094184.

[7] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion for sentiment analysis in twitter," in International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Institute of Advanced Engineering and Science, Oct. 2018, pp. 509–513. doi: 10.1109/EECSI.2018.8752851.

[8] Alvi Rahmy Royyan and Erwin Budi Setiawan, "Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 1, pp. 78–84, Feb. 2022, doi: 10.29207/resti.v6i1.3525.

[9] W. W. Ariestya, I. Astuti, and I. M. Wiryana, "Preprocessing For Crawler Of Short Message Social Media," in 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia: IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/IAC.2018.8780451.

[10] J. Hernandez-Gonzalez, I. Inza, and J. A. Lozano, "A Note on the Behavior of Majority Voting in Multi-Class Domains with Biased Annotators," IEEE Trans Knowl Data Eng, vol. 31, no. 1, pp. 195–200, Jan. 2019, doi: 10.1109/TKDE.2018.2845400.

[11] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," International Journal of Research in Marketing, vol. 36, no. 1, pp. 20–38, Mar. 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[12] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B. W. On, "Fake news stance detection using deep learning architecture (CNN-LSTM)," IEEE Access, vol. 8, pp. 156695–156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[13] M. Anandarajan, C. Hill, and T. Nolan, "Text Preprocessing," 2019, pp. 45–59. doi: 10.1007/978-3-319-95663-3_4.

[14] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Jul. 2020. doi: 10.1088/1757-899X/874/1/012017.

[15] J. Yao, "Automated Sentiment Analysis of Text Data with NLTK," in Journal of Physics: Conference Series, Institute of Physics Publishing, May 2019. doi: 10.1088/1742-6596/1187/5/052020.

[16] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Measuring information credibility in social media using combination of user profile and message content dimensions," International Journal of Electrical and Computer Engineering, vol. 10, no. 4, pp. 3537–3549, 2020, doi: 10.11591/ijece.v10i4.pp3537-3549.

[17] L. Dhara J and D. Nikita P, "Stopword Identification and Removal Techniques on TC and IR Applications: A Survey," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India: IEEE, May 2020. doi: 10.1109/ICACCS48705.2020.9074166.

[18] D. Merlini and M. Rossini, "Text categorization with WEKA: A survey," Machine Learning with Applications, vol. 4, p. 100033, Jun. 2021, doi: 10.1016/j.mlwa.2021.100033.

[19] A. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification Pattern Recognition View project Improvement text classification using log(TF-IDF) with K-NN Algorithm View project," Article in International Journal of Computer Science and Information Security, vol. 16, no. 6, pp. 13–22, 2018, doi: 10.5281/zenodo.1296383.

[20] Zankoya Zaxo and Duhok Polytechnic University, "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF," in 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho - Duhok, Iraq: IEEE, Apr. 2019, pp. 124–128. doi: 10.1109/ICOASE.2019.8723825.

[21] Z. Zhang, Y. Lei, J. Xu, X. Mao, and X. Chang, "TFIDF-FL: Localizing faults using term frequency-inverse document frequency and deep learning," IEICE Trans Inf Syst, vol. E102D, no. 9, pp. 1860–1864, 2019, doi: 10.1587/transinf.2018EDL8237.

[22] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," Int J Comput Appl, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[23] A. Nurdin, B. Anggo, S. Aji, A. Bustamin, and Z. Abidin, "PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS," Jurnal TEKNOKOMPAK, vol. 14, no. 2, p. 74, 2020, doi: https://doi.org/10.33365/jtk.v14i2.732.

[24] E. M. Dharma, F. Lumban Gaol, H. Leslie, H. S. Warnars, and B. Soewito, "THE ACCURACY COMPARISON AMONG WORD2VEC, GLOVE, AND FASTTEXT TOWARDS CONVOLUTION NEURAL NETWORK (CNN) TEXT CLASSIFICATION," J Theor Appl Inf Technol, vol. 31, no. 2, 2022, [Online]. Available: www.jatit.org

[25] L. Deng et al., "News Text Classification Method Based on the GRU_CNN Model," International Transactions on Electrical Energy Systems, vol. 2022, 2022, doi: 10.1155/2022/1197534.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 386−396
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4057

[26] S. Sridevi, G. R. Karpagam, and B. V. Kumar, "GENETIC ALGORITHM - OPTIMIZED GATED RECURRENT UNIT (GRU) NETWORK FOR SEMANTIC WEB SERVICES CLASSIFICATION," Malaysian Journal of Computer Science, vol. 35, no. 1, pp. 70–88, 2022, doi: 10.22452/mjcs.vol35no1.5.

[27] M. A. Hossain, R. Karim, R. Thulasiram, N. D. B. Bruce, and Y. Wang, "Hybrid Deep Learning Model for Stock Price Prediction," in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India: IEEE, 2018, pp. 1837–1844. doi: 10.1109/SSCI.2018.8628641.

[28] 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.

[29] M. M. Fahmy, "Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial," Journal of Engineering Research, vol. 6, no. 5, 2022, doi: 10.21608/ERJENG.2022.274526.