Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Effectiveness of Bi-GRU and FastText in Sentiment Analysis of Shopee App Reviews

Rayhan Fadhil Rahmanda[1)*], Yuliant Sibaroni[2)], Sri Suryani Prasetiyowati[3)]
[1,2,3)]School of Computing, Telkom University, Bandung, Indonesia
[1)]rayhanfr@student.telkomuniversity.ac.id, [2)]yuliant@telkomuniversity.ac.id,
[3)]srisuryani@telkomuniversity.ac.id

Submitted : Jan 1, 2025 | **Accepted** : Feb 1, 2025 | **Published** : Feb 3, 2025

Abstract: E-commerce is proof of evolution in the economic field due to its flexibility to shop for various necessities of life anytime and anywhere. Shopee is one of the e-commerce platforms in demand by people from varied circles in Indonesia. Multiple reviews are shed publicly by Shopee users on the Google Play Store regarding shopping experiences, which can be positive or negative. This condition affects the decision of other users to shop at Shopee, thus impacting the increase or decrease in profits from Shopee itself. Therefore, user sentiment analysis is needed as a form of effort to maintain user trust in Shopee. This research aims to build a system to classify the sentiment of Shopee application users through reviews in the Google Play Store by utilizing the Bidirectional Gated Recurrent Unit (Bi-GRU) deep learning model. The dataset contains 9,716 reviews, including 3,937 positive and 5,779 negative sentiments. Several test scenarios were conducted to achieve the highest peak of performance, utilizing TF-IDF feature extraction, FastText feature expansion, and optimization using the Cuckoo Search Algorithm. Additionally, SMOTE resampling was utilized to correct the dataset's uneven distribution. The combined test scenarios mentioned significantly improved the accuracy by 1.03% and F1-Score by 1.04% from the baseline, with the highest accuracy reaching 90.48% and the highest F1-Score of 90.16%.

Keywords: Bi-GRU, Cuckoo Search Algorithm, FastText, Sentiment Analysis, Shopee

INTRODUCTION

The progress in science and technology has paved the way for the digital transformation of trade, primarily through the rise of e-commerce platforms (Widyastuti & Prastitya, 2020). Through e-commerce, buying and selling transactions of goods and services can be carried out online without distance and time restrictions (Fauzi & Lina, 2021). Shopee ranks as one of the most prominent e-commerce platforms in Indonesia, having surpassed 100 million downloads and receiving nearly 8 million reviews on the Google Play Store (Afdal & Waroka, 2022). Customers often give reviews on the Google Play Store to share their experiences, which can influence other users' shopping decisions (Limbong et al., 2022). For Shopee, these reviews become an essential data source for improving services and strengthening competitiveness through sentiment analysis.

Sentiment analysis, also known as opinion mining, is the process that involves investigating beliefs, feelings, and emotions through calculations. This analysis helps to identify positive or negative sentiments of a text (Ridwansyah, 2022). This widely utilized method is recognized for its significant role in identifying public perceptions regarding a product (Utami, 2022). To enhance the accuracy of the analysis, employing deep learning-based methods stands out as a highly effective approach (Suhartono et al., 2022).

Deep learning is considered more powerful than machine learning in text classification (Gochhait, 2024). One of the deep learning models utilized in sentiment analysis is the Bidirectional Gated Recurrent Unit (Bi-GRU). This model is designed to process word context in both directions, enhancing its effectiveness, especially when integrated with word embedding techniques (Yin et al., 2021). In research (Pratama & Cahyono, 2024) related to public sentiment towards corruption, Bi-GRU combined with Word2Vec achieved the highest accuracy of 88.2%, outperforming RNN, LSTM, GRU, and Bi-LSTM. Despite its advantages, there are opportunities to improve the performance of sentiment analysis, including through feature expansion and the application of optimization techniques.

FastText is a word embedding method that extends the current implementation of Word2Vec. FastText outperforms Word2Vec and GloVe word embedding (Dharma et al., 2022). FastText can be used as a feature

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

expansion tool to overcome vocabulary-matching errors and enrich text representation. In research (Yahya & Setiawan, 2022), FastText was applied to expand TF-IDF features in Gradient Boosted Decision Tree model for Twitter (X) topic classification, successfully yielding an accuracy of 91.39%. In addition, research (Jawad et al., 2023) and (Sudhamathy & Valliammal, 2023) also proved that using the Cuckoo Search Algorithm (CSA) optimization can improve model performance in text classification.

Based on previous research, researchers are motivated to implement Bi-GRU in classifying the sentiment of Shopee app users' reviews in Indonesian, given its advantages over other methods. In addition, TF-IDF feature extraction, FastText feature expansion, and CSA optimization will be tested to see the effectiveness of testing on model performance. Each test scenario is designed as a form of effort to improve model performance and is comprehensively explained in this study.

LITERATURE REVIEW

Research on sentiment analysis utilizing deep learning has been conducted before. Research (Pratama & Cahyono, 2024) compared several deep learning methods, namely RNN, LSTM, GRU, Bi-LSTM, and Bi-GRU, to analyze sentiment towards corruption using English datasets. The dataset is distributed into three sentiment categories: positive, neutral, and negative. By adding Word2Vec word embedding and tuning parameters on the number of units and epochs, Bi-GRU attained its maximum accuracy and F1-Score of 88% and 87.51%, respectively.

Research (Sachin et al., 2020) examined sentiment analysis on a dataset of Amazon user reviews. Three classes of sentiment labels, positive, negative, and neutral, cover 120,000 reviews. The tested models include LSTM, Bi-LSTM, GRU, and Bi-GRU, which are designed using a simple architecture: embedding layer, LSTM/GRU-based layer, and dense layer as an output. With a balanced dataset, utilizing GloVe word embedding, Softmax activation function, and dropout regularization, Bi-GRU yielded an accuracy of 71.19%, followed by GRU with 71.06% accuracy. These outcomes denote that the GRU-based model excels over the performance of the LSTM-based model.

Several studies have applied FastText word embedding for text classification. For example, research (Dharma et al., 2022) incorporated word embedding in a CNN model to classify 19,977 English news articles from the UCCI KDD Archive into 20 distinct categories. The researchers compared the capabilities of three word embeddings, namely GloVe, Word2Vec, and FastText, each with a vector dimension of 300. The results showed that using FastText reached the best accuracy of 97.2%. In addition, FastText also performs effectively as a feature expansion. Research (Yahya & Setiawan, 2022) utilized FastText feature expansion to classify 11 topics on the Indonesian Twitter(X) dataset. FastText was used to expand the features extracted by TF-IDF. The application of FastText in the Gradient Boosted Decision Tree model enhanced accuracy by 0.51% compared to TF-IDF alone.

Research (Novitasari & Purbolaksono, 2021) highlights using TF-IDF and n-gram feature extraction in sentiment analysis based on aspect level in beauty product reviews. This research uses the Naive Bayes model and feature selection of Chi-Square. Combining the two feature extractions in the Naive Bayes model and Chi-Square as feature selection achieved 80.18% accuracy. Furthermore, research (Jawad et al., 2023) utilized deep learning combined with the Cuckoo Search Algorithm (CSA) to detect a sign of depression. This approach was compared with machine learning-based methods from previous studies, and the results showed that deep learning achieved the highest accuracy of 99.5%, surpassing the performance of machine learning.

Based on the above research, the Bi-GRU model has been proven to be superior to other methods in text classification, so it has been selected for this study. The proposed model is further integrated with TF-IDF feature extraction, FastText feature expansion, and CSA optimization to improve the sentiment analysis performance of Shopee app users. In addition, this research also utilizes the Synthetic Minority Oversampling Technique (SMOTE) to balance the sentiment distribution in the dataset used.

METHOD

This research aims to develop a sentiment classification system for Shopee app reviews. Fig. 1 describes all the stages of the research in detail. The main stages include data collection, data labeling, data preprocessing, classification using Bi-GRU, feature extraction with TF-IDF, SMOTE balancing, feature expansion with FastText, Cuckoo Search algorithm optimization, and system performance evaluation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

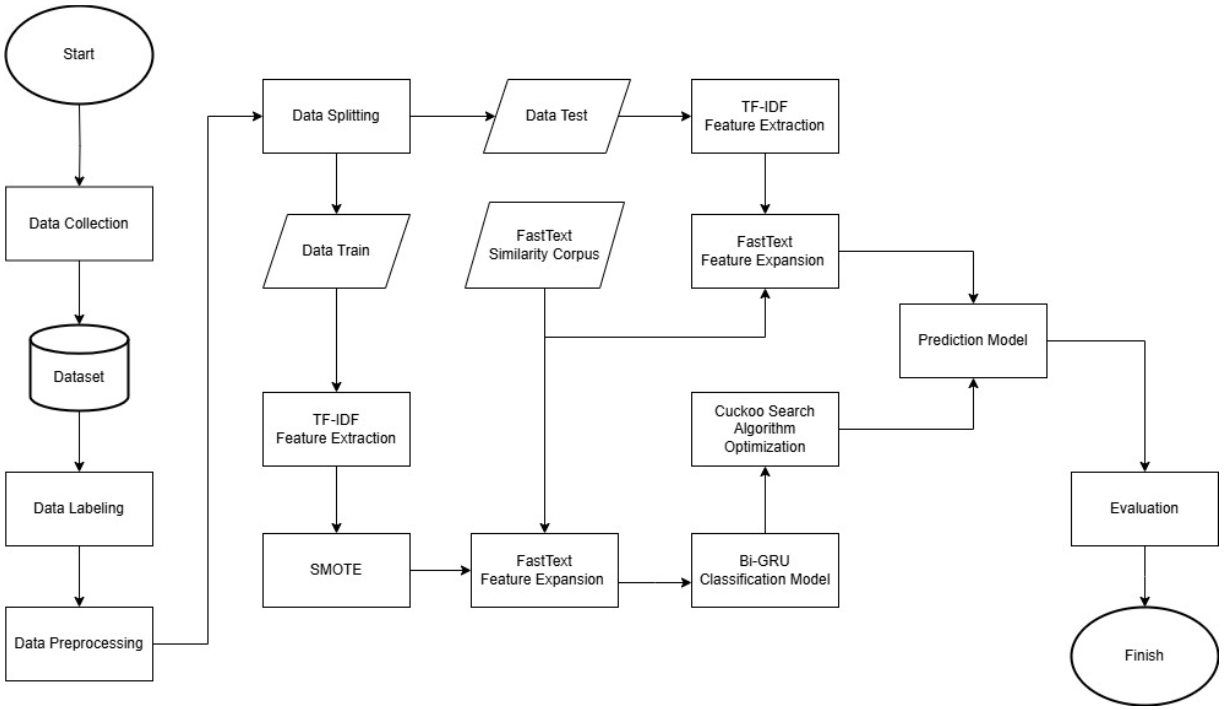

Fig. 1 System Flowchart

Dataset

Data was collected through crawling process of the Shopee app reviews on the Google Play Store using the google-play-scraper library. A total of 10,702 data entries were collected and stored as a dataset in Excel format. After collecting the data, a manual labeling process was conducted to ensure consistency and compatibility between each review and the label given. Labels are divided into positive (1) for reviews expressing satisfaction and praise and negative (0) for reviews expressing disappointment and slurs against the service. Examples of data that have been labeled can be seen in Table 1.

Table 1. Dataset Labeling

Data	Label
saya tidak suka aplikasi ini....alamat saya susah di ubah. nyesal saya dowload aplikasi ini.. buang2 paket saya aja	0
...	...
Jangan ragu pakai Shopee sdh terbukti kualitas barang2nya dan harganya juga murah	1

Data Preprocessing

Data preprocessing solves problems with raw data that may contain errors, discrepancies, or incompleteness so that the resulting data set becomes more useful (Lubis et al., 2021). Preprocessing involves several steps, namely case folding, cleaning, normalization, stopwords removal, stemming, and tokenization.

Case Folding

Case folding is a stage that ensures that the data is consistent by changing the letters in the text that were previously capitalized to small.

Cleaning

Cleaning is a stage of deleting punctuation, symbols, hashtags, numbers, URLs, and other characters that do not affect the overall meaning of each review.

c. Normalization

Normalization is converting slang words, typos, and inappropriate spelling into standard words listed in the Big Indonesian Dictionary (KBBI) and following the Indonesian spelling rules in the Indonesian Spelling Guide (PUEBI).

Stopwords Removal

Stopwords Removal is a stage to delete meaningless words. Examples of removed words are 'pada', 'dan', 'untuk', 'ke', 'dari', 'yang', and others.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Stemming
Stemming is a stage of transforming the words into their initial by deleting their affixes. Affixes can include prefixes, suffixes, or both.

Tokenization
Tokenization is a stage of breaking down sentences contained in the document into individual words. Each word is called a 'token'.

Table 2. Dataset Preprocessing

Preprocessing	Text
Original Review	Aplikasi nya bagus Saya suka ,,, 😄 pokok nya aku suka banget sama aplikasi ini keren gratis ongkir dân bagus pokok nya sukses selalu min !! 👌 😄
Case Folding	aplikasi nya bagus saya suka ,,, 😄 pokok nya aku suka banget sama aplikasi ini keren gratis ongkir dân bagus pokok nya sukses selalu min !! 👌 😄
Cleaning	aplikasi nya bagus saya suka pokok nya aku suka banget sama aplikasi ini keren gratis ongkir bagus pokok nya sukses selalu min
Normalization	aplikasi nya bagus saya suka pokoknya nya aku suka sekali sama aplikasi ini keren gratis ongkos kirim bagus pokoknya nya sukses selalu admin
Stopwords Removal	aplikasi bagus suka pokoknya suka sekali aplikasi keren gratis ongkos kirim bagus pokoknya sukses admin
Stemming	aplikasi bagus suka pokok suka kali aplikasi keren gratis ongkos kirim bagus pokok sukses admin
Tokenization	['aplikasi', 'bagus', 'suka', 'pokok', 'suka', 'kali', 'aplikasi', 'keren', 'gratis', 'ongkos', 'kirim', 'bagus', 'pokok', 'sukses', 'admin']

TF-IDF Feature Extraction
TF-IDF is widely used to calculate various terms on text documents. It combines both TF and IDF concepts. TF (Term Frequency) detects the upward trend of a term's usage and counts it, and IDF (Inverse Document Frequency) reduces the weight of a term if it is widespread in the document (Prabowo & Azizah, 2020). This research uses TF-IDF as a tool for feature extraction. TF-IDF Feature Extraction can be calculated through the following equations: (1), (2), and (3).

$$tf_{t,d} = \sum_{x \in d} f_t(x) \quad (1)$$

$$idf_t = log \frac{N}{df_t} \quad (2)$$

$$tfidf_{t,d} = tf_{t,d} * idf_t \quad (3)$$

SMOTE Balancing
The Synthetic Minority Oversampling Technique (SMOTE) balances the proportion between majority and minority classes in the dataset by resampling the minority classes. This approach not only improves the representation of minority classes but also helps to reduce potential errors in predicting classification results, thus optimizing model performance (Anggraeni & Andryana, 2024).

FastText Feature Expansion
FastText is a word embedding model developed by the artificial intelligence research team at Facebook. The model reportedly has about 200 million vocabularies with 300 dimensions and 600 billion word vectors. Thanks to the simplicity of its architecture, FastText makes it possible to handle text classification jobs quickly and precisely. Its working mechanism focuses on converting words into an n-gram representation of characters (Umer et al., 2023). For example, the word 'bagus' with 3-gram characters can be decomposed into embedding spaces containing word vectors, namely <ba, bag, agu, gus, us>. The word vector 'good' is obtained from the sum of the n-gram character vectors. This illustration shows the advantage of FastText in utilizing subwords for semantic information processing, thus improving the model's ability to handle words outside the vocabulary and positively impacting the model's performance. (Raihan & Setiawan, 2022). Two main architectures in FastText are Skipgram and Continous Bag of Words (CBOW). Both architectures work in opposite ways. The CBOW model predicts the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

target word through inputs derived from the surrounding context words, whereas Skip-Gram utilizes the target word as input to predict the surrounding context words. Fig. 2 presents the architecture of the two models.

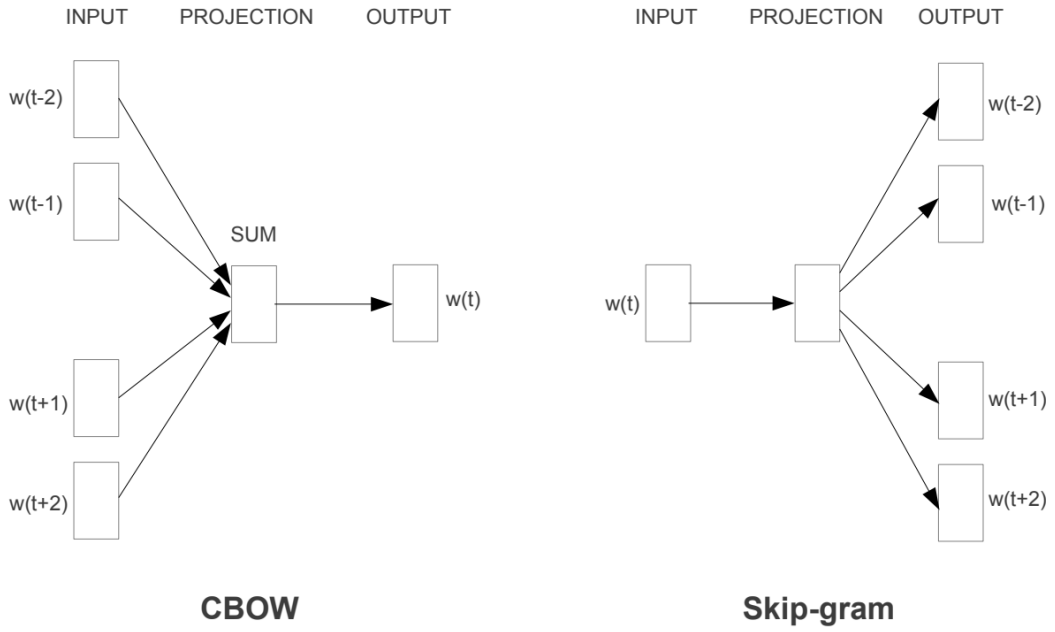

Fig. 2 FastText Model Architecture (Mikolov et al., 2013)

This research uses a pre-trained corpus trained from Wikipedia and Common Crawl with the CBOW approach. The corpus is used to expand features through Top-n similar words available in the corpus (Abasan & Setiawan, 2024). The model can obtain a richer representation and improve its overall performance by applying FastText as a feature expansion tool.

Bidirectional Gated Recurrent Unit (Bi-GRU) Classification

Bidirectional Gated Recurrent Unit (Bi-GRU) is one of the Recurrent Neural Network-based deep learning models that is widely used in various research tasks, such as sentiment analysis. Bi-GRU is an extension of GRU, where Bi-GRU uses two layers of GRU, which makes it able to work in two directions through each of the two GRU inputs, namely the forward and backward directions. This makes Bi-GRU more powerful than GRU in capturing information, as it considers both past and future context, making it suitable for use on sequential data. (Shaikh & Ramadass, 2024). Simultaneously, the respective hidden states of the two input GRUs are combined to calculate the output cell (Meng et al., 2020). Fig. 3 (Faadhilah & Setiawan, 2024) shows the architecture of the Bi-GRU model.

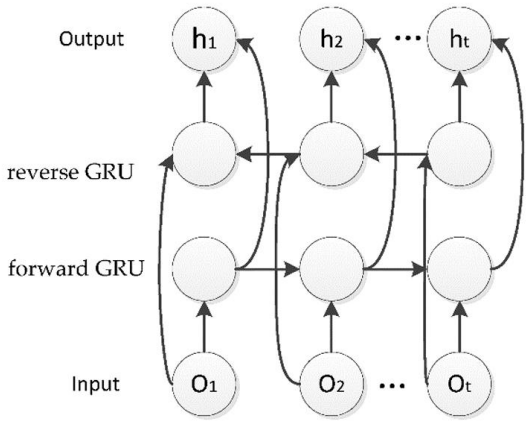

Fig. 3 Model Architecture of Bi-GRU

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

This research was implemented using Google's TensorFlow library and run through Jupyter Notebook using Python. The model architecture consists of one Bi-GRU layer and an activation function on the output layer. In addition, recurrent dropout, dropout regularization, and early stopping mechanisms are used to reduce the risk of overfitting.

Cuckoo Search Algorithm Optimization

The Cuckoo Search Algorithm (CSA) is a metaheuristic search algorithm developed by Yan and Deb in 2009. This algorithm takes the concept from the behavior of cuckoo birds in finding other birds' nests for them to lay their eggs. They rely on nest-owning birds to care for their eggs (Yang & Deb, 2010). In CSA, each cuckoo bird egg is treated as a candidate solution in the optimization process. Candidate solutions are obtained through the Levy Flights strategy, which allows exploring candidate solutions through a global random walk method. If the solution obtained does not meet the probability requirements, the candidate will be discarded and replaced with a new solution candidate (Civicioglu & Besdok, 2013). In the context of this research, CSA is used as an optimization method to find the best hyperparameter values in the Bi-GRU model, with the expectation of the model's ability to reach its maximum potential.

Performance Evaluation

This stage measures the system performance in predicting test data by measuring True Negative (TN), False Negative (FN), True Positive (TP), and False Negative (FN). Table 3. displays the Confusion Matrix of the sentiment label prediction results on the test data. TN refers to negative sentiment that the model correctly predicts as a negative sentiment. TP refers to a positive sentiment that is correctly predicted as a positive sentiment. FP is a negative sentiment mistakenly predicted as a positive sentiment. FN is a positive sentiment that is mistakenly predicted to be a negative sentiment.

Table 3. Confusion Matrix

Actual Label	Predicted Label	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

The measurement of the number of TN, TP, FP, and FN will consider the accuracy, precision, recall, and F1-score values. Detailed information regarding the calculation of model performance evaluation metrics can be observed below:

Accuracy is a division between the total actual positive and negative predictions with the overall predictions generated by the model. The calculation method is shown in equation (4).

$$accuracy = \frac{TN + TP}{TP + FN + TN + FP} \times 100\% \qquad (4)$$

Precision is calculating the number of correct positive predictions compared to the total number of positives. The calculation method is shown in equation (5).

$$precision = \frac{TP}{TP + FP} \times 100\% \qquad (5)$$

Recall is a process for measuring the model's ability by calculating the ratio of the number of positive predictions taken to the total number of positives. The calculation method is shown in equation (6).

$$recall = \frac{TP}{FP + FN} \times 100\% \qquad (6)$$

F1-Score calculates model performance based on the average value of precision and recall. The calculation method is shown in equation (7).

$$f1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (7)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

Dataset Preparation

The initial stage of research is to collect data by crawling the Shopee app reviews available on the Google Play Store. All collected reviews are then saved in Excel format and manually labeled to be classified into positive and negative sentiments. A preprocessing stage is carried out after the labeling process is complete to facilitate the model in the training process. During preprocessing, 986 null data were deleted, leaving 9,716 data for research. The sentiment distribution of the clean dataset can be observed in Fig. 4.

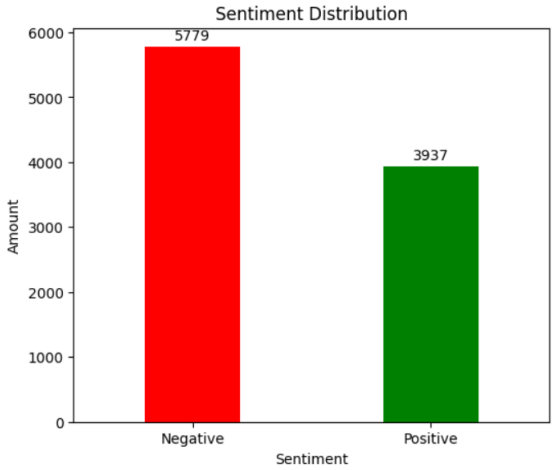

Fig. 4 Sentiment Distribution on Clean Dataset

Bi-GRU Modeling

The modeling process consists of four scenarios to obtain the best performance of the Bi-GRU model. The first scenario establishes the optimal splitting ratio of training and test data, which will be used as the baseline for further testing. The second scenario applies TF-IDF feature extraction from the baseline with SMOTE to handle data imbalance. The third scenario applies FastText feature expansion. The fourth scenario applies the Cuckoo Search Algorithm (CSA) to find the best parameter values.

First Scenario (Baseline)

The first scenario aims to find a baseline model by testing with training and test data splitting ratios of 90:10 and 80:20. Table 4. shows some of the hyperparameters and values used as baselines. This scenario uses the embedding layer because feature extraction has not been implemented.

Table 4. Bi-GRU Parameter

Parameter	Value
embedding size	27
units	64
recurrent_dropout	0.4
dropout	0.4
epochs	10
batch size	20
early stopping	4
optimizer	RMSprop
loss function	binary crossentropy
learning rate	0.001

Table 5. shows that the 80:20 splitting ratio achieved an accuracy of 89.45% and an F1-Score of 89.12%. Therefore, the 80:20 ratio is set as the baseline for the next scenario.

Table 5. First Test Scenario

Split Ratio	Accuracy (%)	F1-Score (%)
90:10	88.58	88.36
80:20	**89.45**	**89.12**

*name of corresponding author

Second Scenario (Baseline + TF-IDF + SMOTE)

In the second scenario, TF-IDF feature extraction is used in the baseline. Based on Fig. 5, the data distribution is not balanced. Therefore, data balancing is performed using SMOTE. The parameter of TF-IDF, namely the max feature, will be tested to select the best max features to be used for the following scenario, with the max feature values to be tested: 500, 2500, 5000, 7500, and 10000. From Table 6., the max feature value of 2500 achieved the highest accuracy of 89.97%, with an increase of 0.52% from the baseline, and the highest F1-Score of 89.64%, with an increase of 0.42% from the baseline. The max feature value of 2500 was chosen for use in the following scenario.

Table 6. Second Test Scenario

Max Feature	Accuracy (%)	F1-Score (%)
500	89.56 (+0.11)	89.22 (+0.10)
2500	**89.97 (+0.52)**	**89.64 (+0.52)**
5000	89.04 (-0.41)	88.65 (-0.47)
7500	89.81 (+0.36)	89.51 (+0.39)
10000	89.87 (+0.42)	89.50 (+0.38)

Third Scenario (Baseline + TF-IDF + SMOTE + FastText)

In the third scenario, the implementation of FastText feature expansion is added. The test uses the pre-trained Wikipedia-Common Crawl corpus to expand the features by testing Top-1, Top-2, Top-3, Top-5, and Top-10 similar words. From Table 7., the implementation of feature expansion with Top-10 similar words achieved the highest accuracy of 90.23%, with an improvement of 0.78% from the baseline, and the highest F1-Score of 89.89 with an improvement of 0.77% from the baseline. The corpus with Top-10 similar words is used for the following scenario.

Table 7. Third Test Scenario

Top (n)	Accuracy (%)	F1-Score (%)
1	89.87 (+0.42)	89.56 (+0.44)
2	89.81 (+0.36)	89.50 (+0.38)
3	89.81 (+0.36)	89.47 (+0.35)
5	90.02 (+0.57)	89.71 (+0.59)
10	**90.23 (+0.78)**	**89.89 (+0.77)**

Fourth Scenario (Baseline + TF-IDF + SMOTE + FastText + CSA)

In the fourth scenario, CSA optimization is used to find the optimal parameter values for Bi-GRU to achieve maximum performance in classifying sentiments. Performance is measured using the best fitness value, with model accuracy as the fitness value. A description of the CSA parameters utilized for model optimization can be seen in Table 8. The selection of Bi-GRU parameters was narrowed down to dropout and learning rate because both tend to improve model performance while keeping the model from overfitting. The results of finding the optimal dropout and learning rate parameter values through CSA optimization are presented in Table 9. Based on these results, the optimal parameters were obtained in the model through CSA optimization, with the best accuracy of 90.48% and F1-Score of 90.16%. The improvement from the baseline is 1.03% for accuracy and 1.04% for F1-Score.

Table 8. CSA Parameter

Parameter	Value	Description
pop_size	40	number of candidate solutions
max_iter	3	maximum iteration of finding the best candidate solution
alpha	0.2	the probability of discarding a candidate solution

Table 9. Fourth Test Scenario

	Values	Accuracy (%)	F1-Score (%)
Default Parameter	dropout = 0.4, learning_rate = 0.001	90.23 (+0.78)	89.89 (+0.77)
Optimal Parameter	**dropout = 0.034100473580419254, learning_rate = 0.007238817592921395**	**90.48 (+1.03)**	**90.16 (+1.04)**

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14474
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

Four test scenarios were carried out based on the research results described above. With the best data splitting ratio to be used as the baseline model, feature extraction with TF-IDF, data balancing with SMOTE, feature expansion with FastText, and finding the best parameters with CSA optimization, it is proven to significantly improve the performance of the Bi-GRU model in training data to provide more accurate predictions of test data. Table 10. shows the best overall metrics obtained from the Confusion Matrix results of the best testing in Fig. 5. Table 11. shows the increase in accuracy and F1-Score in each test scenario measured from the first scenario, with the highest increase in accuracy of 1.03% and F1-Score of 1.04%.

Table 10. Overall Metrics of Best Testing

Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
90.48	90.25	90.08	90.16

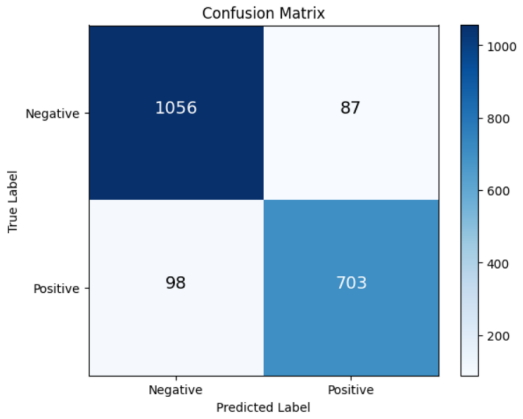

Fig. 5 Confusion Matrix of Best Testing

Fig. 5 shows the best Confusion Matrix of the model prediction results on the test data. This matrix includes True Negative (TN) of 1056, True Positive (TP) of 703, False Negative (FN) of 98, and False Positive (FP) of 87. The high number of TP and TN compared to the low FN and FP indicates that the model is relatively accurate in predicting negative sentiment (TN) and positive sentiment (TP).

Table 11. Performance Increase

Testing Scenario	Accuracy (%)	F1-Score (%)
Baseline	89.45	89.12
Baseline + TF-IDF + SMOTE	89.97 (+0.52)	89.64 (+0.52)
Baseline + TF-IDF + SMOTE + FastText	90.23 (+0.78)	89.89 (+0.77)
Baseline + TF-IDF + SMOTE + FastText + CSA	**90.48 (+1.03)**	**90.16 (+1.04)**

Table 11. illustrates the performance improvement from the four test scenarios performed on the Bi-GRU model, proving the tests' success in improving the model's performance. The first scenario aims to determine the best training and test data separation ratio as a baseline by comparing 90:10 and 80:20 ratios. The results showed that 80:20 ratio gave the best performance with 89.45% accuracy and 89.12% F1-Score. The second scenario adds TF-IDF feature extraction and data balancing using SMOTE to the baseline. The max feature values tested were 500, 2500, 5000, 7500, and 10000. The best results were achieved at max features 2500, with 89.97% accuracy and 89.64% F1-Score, an increase of 0.52% and 0.52%, respectively, from the baseline. The third scenario extends the second scenario by adding FastText feature expansion using pre-trained Wikipedia-Common Crawl corpus. The TF-IDF feature is expanded using similar words (Top-n), with the Top-n values tested including Top-1, Top-2, Top-3, Top-5, and Top-10. The best results were obtained at Top-10 with an accuracy of 90.23% and F1-Score of 89.89%, improving 0.78% and 0.77%, respectively, from the baseline. The fourth scenario applied the CSA optimization technique to find the best parameters of the Bi-GRU model, namely dropout and learning rate. The optimization resulted in a dropout value of 0.034100473580419254 and a learning rate of 0.007238817592921395, which resulted in an accuracy of 90.48% and an F1-Score of 90.16%, an increase of 1.03% and 1.04%, respectively, from the baseline.

This research produces better performance than previous research (Utami, 2022), which used a Recurrent Neural Network (RNN) and a SMOTE-Tomek Links combination to analyze the sentiment of Shopee app reviews

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

in Indonesian, with an accuracy of 80% and an F1-Score of 88.1%. This study's accuracy reached 90.48% and F1-Score 90.16%, significantly improving. These results prove that using Bi-GRU, TF-IDF feature extraction, data balancing with SMOTE, FastText feature expansion, and CSA optimization is an effective combination for sentiment analysis systems.

CONCLUSION

This research successfully adopts the implementation of feature extraction with TF-IDF, feature expansion with FastText, Cuckoo Search Algorithm optimization on the Bi-GRU model, as well as dataset resampling technique with SMOTE in classifying the sentiment of Shopee app user reviews on Google Play Store. The dataset contains 9,716 reviews categorized into positive and negative sentiments. The highest accuracy and F1-Score from the combination of test scenarios previously described were recorded at 90.48% and 90.16%. These values show an increase of 1.03% accuracy and 1.04% F1-Score compared to the first scenario. As suggestions for future research, it is recommended to use more extensive and more complex datasets, apply more sophisticated data preprocessing, try other text representations (e.g., BERT), and explore other deep learning models such as CNN.

REFERENCES

Abasan, I. G. B. J., & Setiawan, E. B. (2024). Empowering hate speech detection: leveraging linguistic richness and deep learning. *Bulletin of Electrical Engineering and Informatics*, *13*(2), 1371–1382. https://doi.org/10.11591/eei.v13i2.6938

Afdal, M., & Waroka, L. (2022). IJIRSE: Indonesian Journal of Informatic Research and Software Engineering Shopee Application Review Classification Using Probabilistic Neural Network Algorithm And K-Nearest Neighbor Klasifikasi Ulasan Aplikasi Shopee Menggunakan Algoritma Probabilistic Neural Network Dan K-Nearest Neighbor. *IJIRSE: Indonesian Journal of Informatic Research and Software Engineering*, *2*(1), 49–58. https://doi.org/10.57152/ijirse.v2i1.216

Anggraeni, S. S., & Andryana, S. (2024). Enhancing Sentiment Analysis Accuracy In Delivery Services Through Synthetic Minority Oversampling Technique. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 257–262. https://doi.org/10.33480/jitk.v9i2.5058

Civicioglu, P., & Besdok, E. (2013). A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. *Artificial Intelligence Review*, *39*(4), 315–346. https://doi.org/10.1007/s10462-011-9276-0

Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. (2022). The Accuracy Comparison Among Word2Vec, GloVe, and FastText Towards Convolutional Neural Network (CNN) Text Classification. *Journal of Theoretical and Applied Information Technology*, *100*(2), 349–359. https://www.jatit.org/volumes/Vol100No2/5Vol100No2.pdf

Faadhilah, A. N., & Setiawan, E. B. (2024). Content-Based Filtering in Recommendation Systems Culinary Tourism Based on Twitter (X) Using Bidirectional Gated Recurrent Unit (Bi-GRU). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, *10*(2), 406–418. https://doi.org/10.26555/jiteki.v10i2.29010

Fauzi, S., & Lina, L. F. (2021). Peran Foto Produk, Online Customer Review dan Online Customer Rating Pada Minat Beli Konsumen di E-Commerce. *Jurnal Muhammadiyah Manajemen Bisnis*, *2*(1), 21. https://doi.org/10.24853/jmmb.2.1.151-156

Gochhait, D. S. (2024). Comparative Analysis of Machine and Deep Learning Techniques for Text Classification with Emphasis on Data Preprocessing. *Qeios*. https://doi.org/10.32388/xhc9j1

Jawad, K., Mahto, R., Das, A., Ahmed, S. U., Aziz, R. M., & Kumar, P. (2023). Novel Cuckoo Search-Based Metaheuristic Approach for Deep Learning Prediction of Depression. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095322

Limbong, J. J. A., Sembiring, I., & Hartomo, K. D. (2022). Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *9*(2), 347–356. https://doi.org/10.25126/jtiik.202294960

Lubis, A. R., Nasution, M. K. M., Sitompul, O. S., & Zamzami, E. M. (2021). The effect of the TF-IDF algorithm in times series in forecasting word on social media. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), 976–984. https://doi.org/10.11591/ijeecs.v22.i2.pp976-984

Meng, S., Jiang, X. Q., Gao, Y., Hai, H., & Hou, J. (2020). Performance Evaluation of Channel Decoder based on Recurrent Neural Network. *Journal of Physics: Conference Series*, *1438*(1). https://doi.org/10.1088/1742-6596/1438/1/012001

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Novitasari, F., & Purbolaksono, M. D. (2021). Sentiment Analysis Aspect Level on Beauty Product Reviews Using Chi-Square and Naïve Bayes. *Journal of Data Science and Its Applications*, *4*(1), 18–030. https://doi.org/10.34818/JDSA.2021.4.72

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Prabowo, W. A., & Azizah, F. (2020). Sentiment Analysis for Detecting Cyberbullying Using TF-IDF and SVM. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(6). https://doi.org/10.29207/resti.v4i6.2753

Pratama, R. G. A., & Cahyono, N. (2024). Comparison of Deep Learning Methods on Sentiment Analysis Using Word Embedding. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(1), 1–8. https://doi.org/10.33480/jitk.v10i1.5280

Raihan, M. A., & Setiawan, E. B. (2022). Aspect Based Sentiment Analysis with FastText Feature Expansion and Support Vector Machine Method on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(4), 591–598. https://doi.org/10.29207/resti.v6i4.4187

Ridwansyah, T. (2022). KLIK: Kajian Ilmiah Informatika dan Komputer Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier. *Media Online*, *2*(5), 178–185. https://djournals.com/klik

Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment Analysis Using Gated Recurrent Neural Networks. In *SN Computer Science* (Vol. 1, Issue 2). Springer. https://doi.org/10.1007/s42979-020-0076-y

Shaikh, Z. M., & Ramadass, S. (2024). Unveiling deep learning powers: LSTM, BiLSTM, GRU, BiGRU, RNN comparison. *Indonesian Journal of Electrical Engineering and Computer Science*, *35*(1), 263–273. https://doi.org/10.11591/ijeecs.v35.i1.pp263-273

Sudhamathy, G., & Valliammal, N. (2023). Twitter Sentiment Analysis Using Clustering based Cuckoo Search Algorithm and Ensemble Classifier. *Proceedings of the 2023 International Conference on Intelligent Systems for Communication, IoT and Security, ICISCoIS 2023*, 212–217. https://doi.org/10.1109/ICISCoIS56541.2023.10100606

Suhartono, D., Purwandari, K., Jeremy, N. H., Philip, S., Arisaputra, P., & Parmonangan, I. H. (2022). Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews. *Procedia Computer Science*, *216*, 664–671. https://doi.org/10.1016/j.procs.2022.12.182

Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. *Multimedia Tools and Applications*, *82*(4), 5569–5585. https://doi.org/10.1007/s11042-022-13459-x

Utami, H. (2022). Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network. *Indonesian Journal of Applied Statistics*, *5*(1), 31. https://doi.org/10.13057/ijas.v5i1.56825

Widyastuti, H., & Prastitya, T. A. (2020). Preferensi Konsumen Pengguna E-Commerce yang Memengaruhi Kesadaran akan Perlindungan Konsumen pada Generasi X. *Jurnal Sistem Informasi Bisnis*, *10*(1), 10–19. https://doi.org/10.21456/vol10iss1pp10-19

Yahya, R. A., & Setiawan, E. B. (2022). 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 2022*, 322–327. https://doi.org/10.1109/ICoICT55009.2022.9914896

Yang, X.-S., & Deb, S. (2010). *Engineering Optimisation by Cuckoo Search*. http://arxiv.org/abs/1005.2908

Yin, X., Liu, C., & Fang, X. (2021). Sentiment analysis based on BiGRU information enhancement. *Journal of Physics: Conference Series*, *1748*(3). https://doi.org/10.1088/1742-6596/1748/3/032054

*name of corresponding author