ISSN 2723-6471

Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset

M. Khairul Anam[1,*], Munawir[2], Lusiana Efrizoni[3], Nurul Fadillah[4], Wirta Agustin[5], Irwanda Syahputra[6], Tri Putri Lestari[7], Muhammad Bambang Firdaus[8], Lathifah[9], Atalya Kurnia Sari[10]

[1,2,4,6]*Universitas Samudra, Jl, Prof. Dr. Syarief Thayeb, Meurandeh, Langsa and 24416, Indonesia*

[3,5]*Universitas Sains dan Teknologi Indonesia, Jl. Purwodadi Indah, Km. 10 Panam, Pekanbaru and 28294, Indonesia*

[7,10]*Universitas Indraprasta PGRI, Jl. Nangka Raya, Jakarta and 12530, Indonesia*

[8]*Universitas Mulawarman, Jl. Ahmad Yani, Batam and 29461, Indonesia*

[9]*Universitas Teknokrat Indonesia, Jl. ZA. Pagar Alam, Bandar Lampung and 35132, Indonesia*

(Received: September 21, 2024; Revised: October 13, 2024; Accepted: November 16, 2024; Available online: December 29, 2024)

Abstract

This study addresses emotion detection challenges in tweets, focusing on contextual understanding and class imbalance. A novel hybrid deep learning architecture combining GRU-BiLSTM with SMOTE is proposed to enhance classification performance on an Israel-Palestine conflict dataset. The dataset contains 40,000 tweets labeled with six emotions: anger, disgust, fear, joy, sadness, and surprise. SMOTE effectively balances the dataset, improving model fairness in detecting minority classes. Experimental results show that the GRU-BiLSTM hybrid with an 80:20 data split achieves the highest accuracy of 89%, surpassing BiLSTM alone, which obtained 88%, and other state-of-the-art models. Notably, the proposed model delivers significant improvement in detecting the emotion of joy (recall: 0.87, F1-score: 0.86). In contrast, the surprise category remains challenging (recall: 0.24). Compared to existing research, this study highlights the effectiveness of combining SMOTE and hybrid GRU-BiLSTM, outperforming models such as CNN, GRU, and LSTM on similar datasets. The incorporation of GloVe embeddings enhances contextual word representations, enabling nuanced emotion detection even in sarcastic or ambiguous texts. The novelty lies in addressing class imbalance systematically with SMOTE and leveraging GRU-BiLSTM's complementary strengths, yielding superior performance metrics. This approach contributes to advancing emotion detection tasks, especially in conflict-related social media data, by offering a robust, context-sensitive, and balanced classification method.

Keywords: BiLSTM-GRU, Emotion, GloVe, SMOTE, Twitter

1. Introduction

Israeli-Palestinian conflict has a complex background involving history, politics, religion, and society [1]. Starting from the Zionist movement by Theodor Herzl at the end of the 19th century, which aimed to establish a Jewish state in Palestine, causing tensions with the local Arab population [2]. After World War I, the region came under British mandate, and tensions increased with large Jewish immigration. In 1947, the UN proposed dividing Palestine into two states, but this plan was rejected by Arab countries, triggering war in 1948 after Israel declared independence [3]. This conflict continues to this day, fueled by issues such as the status of Jerusalem, borders, the rights of Palestinian refugees, and Israeli settlements in the occupied territories [4].

The public response on social media to the Israeli-Palestinian conflict has been varied and intense, reflecting a wide range of views and emotions from users around the world. Many social media users show solidarity with one side through hashtag campaigns, profile pictures, and sharing information or misinformation about the conflict. Hashtags such as #FreePalestine and #StandWithIsrael frequently trend, reflecting support or protest from various global communities. Vigorous discussions and debates often occur on platforms such as Twitter, Facebook, and Instagram,

***Corresponding author: M. Khairul Anam (khairulanam@unsam.ac.id)**
DOI: https://doi.org/10.47738/jads.v6i1.459

with many users sharing their political, historical, and humanitarian views. In addition, many public figures, celebrities, and humanitarian organizations participated in these discussions to voice their opinions and raise awareness about the situation. Videos, images, and stories from conflict victims often go viral, stirring netizens' emotions and actions of solidarity. However, social media is also a place where misinformation and propaganda can spread, fueling further tensions and exacerbating conflicts. Therefore, public responses on social media are very significant in shaping global opinions and perceptions about the Israeli-Palestinian conflict.

The public's emotional response on social media to the Israeli-Palestinian conflict includes various emotions such as joy, sadness, anger, fear, disgust, and surprise. The emotion of joy arises when there is good news, such as a ceasefire or peace agreement, which gives hope to many people. In contrast, sadness is a very common response to news about casualties, injuries, and destruction, especially images and videos that show human suffering. Anger is often seen in reactions to violence and injustice felt by both parties. Fear arises from the uncertainty and constant threat felt by those involved in or affected by conflict. Disgust is a response to actions that are considered inhumane or brutal. Lastly, suppression reflects an individual's attempt to suppress or avoid excessive emotional involvement, perhaps due to exhaustion or hopelessness in a seemingly never-ending situation.

On social media, emotions like joy can arise from various contexts, including positive events such as ceasefires. However, the complexity of tweets that may contain sarcasm or irony can pose challenges for emotion classification models, especially in detecting genuine joy. For example, if a user implies distrust or skepticism, it may be closer to disgust. Meanwhile, if the statement contains frustration about a situation, such as dissatisfaction with the outcome of a ceasefire, it would be more accurately classified as anger. This approach helps reduce ambiguity in emotion classification by providing a deeper context for tweets containing sarcasm or irony, ensuring that emotion classification is not solely based on positive or negative words but also on a broader understanding of the emotional context.

Previous research has discussed emotion classification using deep learning algorithms. Research conducted by [5] carried out sentiment analysis using Attention based BiLSTM and obtained an accuracy of 79.68%. Other researchers used ERNIE-BiLSTM to analyze comments, obtaining an accuracy of 88.9% [6]. Furthermore, [7] carried out sentiment analysis using BiLSTM to get an accuracy of 85%. Many current studies also hybridize BiLSTM with other deep learning algorithms. Hybrid Global Vectors for Word Representation (Glove)-Convolutional Neural Network (CNN)-BiLSTM was used for sentiment analysis to get an accuracy of 95.60% [8]. Then hybrid BERT-BiLSTM was used for sentiment analysis, and an accuracy of 93.79% was obtained [9].

However, accuracy often decreases when deep learning analyzes datasets with labels out of 5. For example, BiLSTM with multi-head Attention is used to provide emotional sentiment in comments, getting the highest accuracy of only 76.77% [10]. Then, another person carried out emotional sentiment using CNN-BiLSTM and got an accuracy of 85%. The next research only used a single algorithm, namely BiLSTM, for emotional sentiment and obtained an accuracy of 74.55% [11]. Several studies have been carried out, and research will increase accuracy in analyzing emotional sentiment on social media.

This research also uses deep learning to detect emotions on Twitter concerning Palestine-Israel. Deep learning is used for text-based emotion analysis because of its high ability to understand the context and nuances of natural language, handle unstructured data, and automate complex feature extraction [12]. Models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) can capture sequences and dependencies in text, making them more effective in identifying emotions contained in sentences or paragraphs than traditional methods [13]. BiLSTM is used to identify emotions in text because of its ability to capture context from two directions (forward and backward), allowing a deeper and more accurate understanding of the text's meaning and emotional nuances [14]. BiLSTM is more effective in identifying emotions than one-way models. Then, this research also carried out a GRU-BiLSTM hybrid. This hybrid was used to identify emotions because it combines the strengths of both models, BiLSTM, which captures context from two directions, and GRU, which is simpler and more efficient in computing [15]. This combination increases accuracy and efficiency in understanding and identifying emotions in text.

Using BiLSTM and hybrid GRU-BiLSTM to identify emotions in the text offers significant advantages in context understanding and computational efficiency. BiLSTM captures context from both directions, improving accuracy in recognizing emotions. The GRU-BiLSTM combination combines the strengths of both models for more optimal

results. In addition, applying Synthetic Minority Oversampling Technique (SMOTE) helps overcome the class imbalance problem in the data, ensuring that the model can learn from all classes effectively and provide more accurate and fairer predictions [16].

2. Literature Review

The use of deep learning algorithms for text classification, particularly in sentiment analysis and opinion mining, has garnered significant attention in recent years. Various models such as BiLSTM, GRU, LSTM, and CNN have been widely adopted, each demonstrating strengths in different contexts. Below is a review of previous studies that utilized these algorithms, along with recent advances in combining these models for emotion detection using SMOTE and GloVe. Research conducted by [17] focused on analyzing opinions related to Permendikbud using the BiLSTM algorithm. This study achieved an accuracy of 87% by employing ADAM and RMSprop optimizers, with 25 epochs. BiLSTM, known for its ability to process data bidirectionally (forward and backward), is well-suited for capturing contextual relationships in text, particularly in sentiment analysis. However, while this study demonstrated high accuracy, it did not address challenges such as class imbalance or the use of advanced regularization techniques like dropout to avoid overfitting. Further exploration of these aspects could improve model generalization.

In another related study, [18] utilized the Gated Recurrent Unit (GRU) model to perform text classification, achieving an accuracy of 77%. This study used GloVe for word embeddings to enhance word representation. GloVe is a matrix-based learning method that captures global statistical relationships between words, providing richer word representations. GRU is often chosen for its efficiency over LSTM, as it uses fewer gates while still retaining the ability to capture temporal dependencies in text data. However, the lower accuracy compared to BiLSTM suggests that GRU may not always be the optimal choice for all text classification tasks, particularly in datasets with complex contextual dependencies. Research by [19] applied the LSTM algorithm to analyze sentiment in user reviews of Google Play Store applications. The LSTM model achieved an accuracy of 85%, which reflects its ability to capture sequential patterns in data. LSTM's effectiveness in long-range dependency problems is well-established, making it a suitable choice for sentiment analysis. However, like other RNN-based models, LSTM tends to have longer training times and could benefit from optimizations such as early stopping or hyperparameter tuning, such as adjusting the learning rate.

Additionally, CNNs have been explored for text classification tasks. In the medical domain, [20] applied CNN to classify medical data and achieved an accuracy of 87%. CNNs, traditionally used in image processing, have been adapted for text classification due to their ability to capture local features through convolutional filters. The results of this study show that CNNs can be effectively applied in the medical field, particularly for feature extraction in structured datasets. However, CNNs may be less effective when handling long-range dependencies in text data compared to recurrent models like LSTM or BiLSTM.

In this literature review, various previous studies have examined the effectiveness of models such as BiLSTM, GRU, CNN, and LSTM in natural language processing tasks. However, these studies have not yet been critically compared in terms of methodologies and results. For instance, the BiLSTM model often shows higher accuracy due to its bidirectional data processing capability compared to GRU, which, while efficient, may be less effective in capturing long-term dependencies. This study aims to highlight these differences to identify gaps that previous research has not addressed, including the potential influence of temporal context on the classification accuracy in ongoing events or conflicts.

Furthermore, previous research often overlooks the class imbalance commonly present in social media datasets. Only a few recent studies attempt to address this issue using techniques such as SMOTE, which provides better representation for minority classes and improves model performance on metrics like accuracy, precision, and F1 score [21]. Emphasis on this aspect in our research is expected to fill an existing gap in the literature related to the dynamic nature of social media content. Some studies have also combined models like GRU and BiLSTM to form hybrid models, generally showing improved performance over single models. For instance, combining the efficiency of GRU with the contextual sensitivity of BiLSTM leads to better performance in sentiment analysis on complex datasets such as Twitter. Additionally, embedding techniques like GloVe and BERT are compared in this literature. GloVe, which uses a global

statistical approach, differs from BERT's transformer architecture and may provide deeper contextual embeddings, although it requires greater computational resources.

This research develops a hybrid model that combines GRU-BiLSTM with class balancing techniques like SMOTE to enhance performance in classifying social media text data, specifically on Twitter datasets. Unlike previous studies, which generally use single models or simple hybrids without considering class imbalance, this research focuses on improving model accuracy and contextual sensitivity while addressing class imbalance challenges comprehensively. This approach is expected to achieve superior performance in accuracy, precision, and F1 metrics, reflecting the model's ability to classify minority classes more effectively. The embedding technique used in this study is GloVe, chosen for its ability to capture global contextual representations of words [22]. Unlike other approaches, such as Word2Vec, which uses a skip-gram model to capture semantic relationships [23], FastText, which considers sub-words for flexibility [24], or BERT, which relies on a bidirectional transformer architecture for deeper context [25], this study emphasizes the use of GloVe, which is more suitable for short and informative text data like tweets [26].

3. Methodology

The following is figure 1 which is the flow of the research methodology used to make it easier to carry out trials.

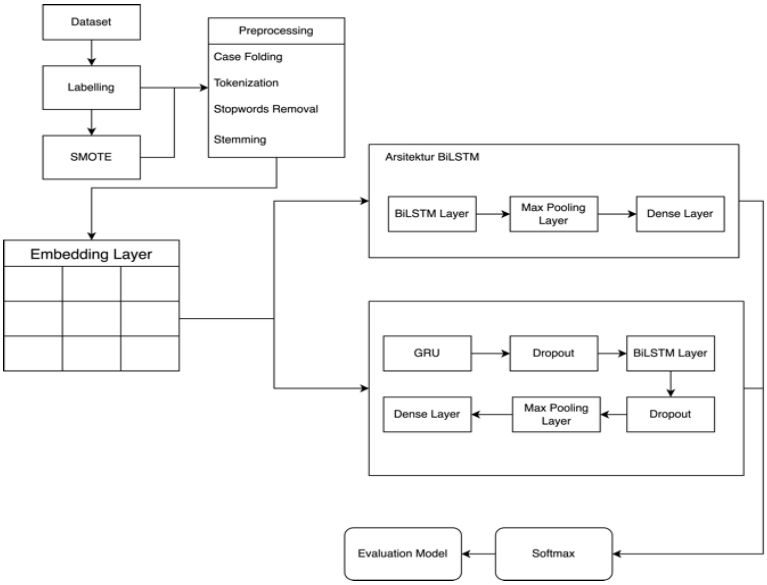


Figure 1. Methodology Flow

3.1.Dataset

The dataset used in this research consists of 40,000 tweets collected from various languages worldwide, with the dataset divided into eight files, each containing 5000 tweets. This data was obtained from the Drone Empirit Academic (DEA) platform using the keyword "Israel-Palestine" and includes tweets collected up to October 26, 2023. At the start of the collection, the dataset had columns such as No, Type, Mentions, Date, Link, Media, Sentiment, Author, Followers, Comments, Likes, Shares, Retweeted, Replied, and Favorited. However, this research only uses the Mentions column and deletes several data records because they have similarities. Tweets that are deleted are mostly retweets because retweets re-spread other people's tweets [27].

To handle the multilingual nature of the dataset, only tweets in English and Indonesian were retained for analysis, while tweets in other languages were excluded to reduce potential biases and ensure consistency. Tweets containing mixed languages (Indonesian and English) were also included, with labeling based on the main context of the tweet. This approach minimizes potential misinterpretations due to linguistic and cultural differences in multilingual texts, allowing for a more focused and accurate analysis, although it limits the generalizability of the results to tweets in these two languages.

 ISSN 2723-6471

3.2. Labelling and Class Balancing with SMOTE

The dataset labeling process uses the NRC Lexicon Emotion library. The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [28]. The dataset is labeled with six basic emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise. Figure 2 is the distribution of comments based on emotion. Figure 2 shows that the data needs to be more balanced; unbalanced data often makes incorrect class predictions on new data. For this reason, it is necessary to balance the data with oversampling techniques using SMOTE. Figure 3 is the result of data balancing using SMOTE.

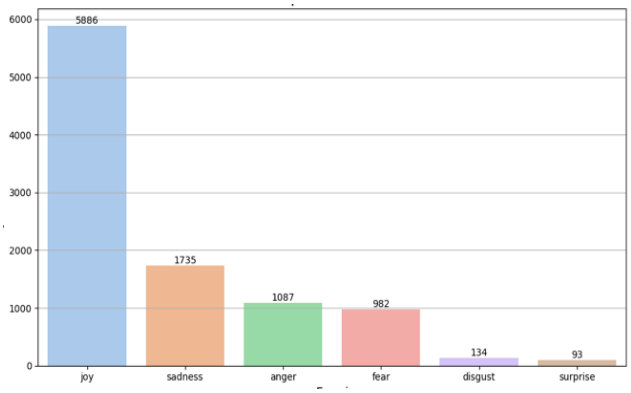


Figure 2. Comment Distribution Based on Emotion

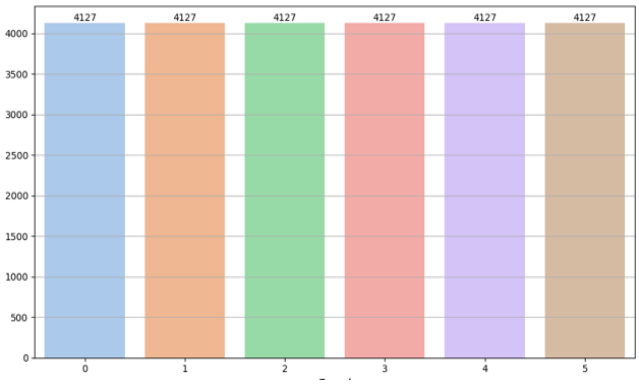


Figure 3. SMOTE-balanced dataset

In this study, the initial dataset exhibited significant class imbalance, with minority classes having substantially fewer data points compared to the majority class. This imbalance affected the initial model's performance, which tended to be more accurate in classifying the majority class while struggling to recognize minority classes, resulting in low precision and F1 scores for those classes. To address this issue, the SMOTE was applied.

Specifically, SMOTE was implemented with a sampling strategy set to match the minority class samples to 50% of the majority class samples. The k-nearest neighbors' parameter (k) was set to 5, which is a common choice to ensure robust synthetic sample generation without excessive overlap. This configuration was selected after initial experimentation to optimize balance and improve minority class representation. By using SMOTE, the data in the minority classes were synthetically increased to achieve a relative balance with the majority class. This step is expected to enhance the overall accuracy of the model as well as improve performance on the minority classes, which initially showed suboptimal results.

3.3. Preprocessing

Next, the text data goes through a preprocessing stage that consists of several steps. Case Folding converts all letters to lowercase for consistency [29]. Tokenization breaks down text into tokens or individual words [30]. Stop word removal removes common words that do not carry much information [31]. Moreover, Stemming converts words to their base forms [32]. This stage is important for cleaning and preparing text data before entering it into the model.

3.4. Glove (Global Vectors for Word Representation)

This research uses the Glove embedding technique to convert tweet text into a vector representation that a machine learning model can process. The Glove was chosen because it can produce vector representations that capture the semantic meaning of words well by utilizing global statistical information from the entire text corpus [33]. Words with similar contexts will have vectors close to each other in the vector space, which is very important in sentiment analysis tasks. Additionally, Glove considers the co-occurrence statistics of words, allowing the model to understand the relationships between words better, which is crucial for determining sentiment in text [34]. Glove is also flexible and can be applied to text in various languages. It is suitable for tweet datasets collected from various languages worldwide, thus enabling consistent and meaningful representation of words in sentiment analysis [8].

 ISSN 2723-6471

3.5. BiLSTM (Bidirectional Long Short-Term Memory)

BiLSTM is a deep learning architecture that processes data in two directions: forward (past to future) and backward (future to past). This bidirectional approach allows BiLSTM to capture context from both preceding and succeeding words in a sequence, enabling a deeper understanding of textual relationships. As depicted in figure 1, BiLSTM serves as one of the core components of the proposed methodology, enhancing the model's ability to detect emotions in text data. In this study, the BiLSTM layer processes text sequences after initial preprocessing and GloVe embedding. GloVe converts words into dense vector representations, which are fed into the BiLSTM layer to capture temporal and contextual dependencies. Following the BiLSTM layer, a Max Pooling operation reduces the dimensionality of the output, highlighting the most significant features for further analysis. Finally, a fully connected Dense layer generates initial predictions based on the extracted features. By utilizing BiLSTM, the model gains the ability to understand complex textual contexts, such as sarcasm or irony, which are often challenging in emotion detection tasks. This architecture's strength lies in its ability to process information from both directions, making it an integral part of the hybrid GRU-BiLSTM approach employed in this research.

3.6. GRU-BiLSTM

The second architecture combines GRU and BiLSTM to improve efficiency and context understanding. The GRU Layer processes text sequences with higher efficiency, and then Dropout is applied to prevent overfitting by randomly ignoring some neurons during training [35]. The BiLSTM Layer processes the text sequence from two directions after GRU, and the Max Pooling Layer reduces the output dimensions to capture important features [36]. A fully connected Dense Layer produces the final prediction.

3.7. Output

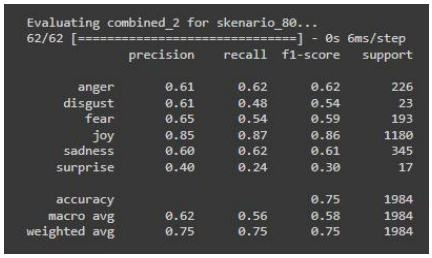
Softmax is used in the output layer to generate class probabilities for each sentiment, and the model is evaluated using confusion metrics consisting of accuracy, precision, recall, and F1-score to assess the sentiment prediction performance. These metrics are chosen because they provide a comprehensive picture of the model's effectiveness in identifying each emotion category, including minority classes that may be more difficult to recognize. Precision and recall are important in this context because they measure the model's ability to avoid misclassification and detect the true emotions present in the data, while F1-score provides a balance between the two metrics. These steps demonstrate how the hybrid GRU-BiLSTM process, with the help of the SMOTE technique, can be used to identify sentiments in text more accurately and efficiently [32].

4. Results and Discussion

To get better results, the dataset needs to be preprocessed before the process is carried out. After that, word weighting was carried out using Glove. Then, it was processed using deep learning.

4.1. Result

Figure 4 displays evaluation metrics such as precision, recall, and F1-score for each emotion category: anger, disgust, fear, joy, sadness, and surprise. The highest recall score is achieved for the emotion category "joy" with a score of 0.87, indicating that the model performs best in detecting the emotion of joy compared to other emotions. In contrast, the lowest recall score is found in the "surprise" category, with a score of 0.24, indicating that the model struggles to recognize the emotion of surprise. On the other hand, the highest F1-score is also achieved for "joy" with a score of 0.86, showing the best balance between precision and recall for this emotion.



```
Evaluating combined_2 for skenario_80...
62/62 [==============================] - 0s 6ms/step
              precision    recall  f1-score   support

       anger       0.61      0.62      0.62       226
     disgust       0.61      0.48      0.54        23
        fear       0.65      0.54      0.59       193
         joy       0.85      0.87      0.86      1180
     sadness       0.60      0.62      0.61       345
    surprise       0.40      0.24      0.30        17

    accuracy                           0.75      1984
   macro avg       0.62      0.56      0.58      1984
weighted avg       0.75      0.75      0.75      1984
```

Figure 4. Classification Report

 ISSN 2723-6471

The average evaluation metrics are also presented at the bottom of the table, with "macro avg" as the average that calculates performance for each class independently, and "weighted avg," which takes into account the data proportion in each class. The overall accuracy of the model is 0.75, indicating fairly good performance, although there are weaknesses in recognizing certain emotions, especially "surprise," which has a low recall score. Figure 5 shows the confusion matrix result from this test.

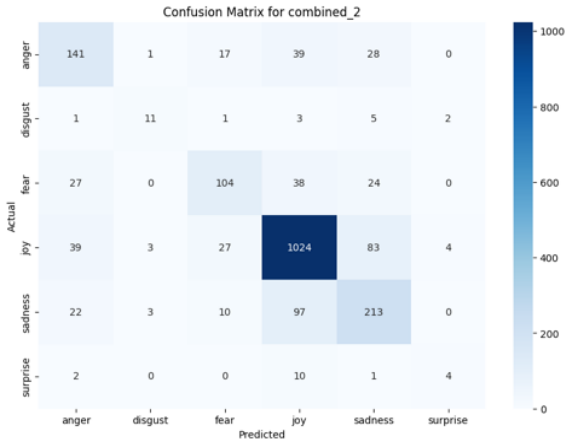


Figure 5. Confusion Matrix

Figure 5 shows the confusion matrix from the results of emotion classification using the GRU-BiLSTM hybrid model. This matrix helps evaluate the model's performance by displaying the number of correct and incorrect predictions for each emotion category. For the "anger" category, the model managed to identify 141 cases correctly but also incorrectly classified several cases as "joy" (39) and "sadness" (28). In the "disgust" category, there were 11 correct predictions, but many errors occurred with the model classifying 11 cases as "fear." For the emotion "fear," the model produced 104 correct predictions, but 27 cases were incorrectly classified as "joy" and 38 as "sadness."

The "joy" category performed best with 1024 correct predictions, although there were some errors where 93 cases were incorrectly classified as "fear" and 83 as "sadness." In the "sadness" category, the model produced 213 correct predictions, but there was a significant error with 97 cases classified as "joy." Finally, for the "surprise" category, correct predictions were only 4, with many errors where cases were classified to various other emotions. Overall, this matrix indicates that the GRU-BiLSTM model performs very well classifying "joy" and "sadness." However, there are challenges in classifying the emotions "disgust" and "surprise" with lower accuracy.

Table 1 present the research results comparing the performance of various models and data split scenarios. Each model was tested with three data split scenarios: 90:10, 80:20, and 70:30, displayed sequentially. The evaluation results include accuracy, precision, recall, and F1-score metrics, compared across models within each data split scenario. This approach ensures that readers can easily follow the performance of each model according to the data conditions used. The models tested are BiLSTM and GRU-BiLSTM, both with and without SMOTE. All models were tested using 30 epochs.

Table 1. Testing Deep Learning Models without SMOTE

No	Model	Splitting Data	without SMOTE				with SMOTE			
			Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
1	BiLSTM	70:30	74%	74%	76%	74%	80%	80%	82%	80%
2		80:20	76%	76%	76%	76%	86%	86%	86%	86%
3		90:10	78%	77%	78%	77%	88%	87%	88%	87%
4	GRU-BiLSTM	70:30	73%	74%	73%	73%	83%	84%	83%	83%
5		80:20	75%	75%	75%	75%	89%	89%	89%	89%
6		90:10	76%	75%	76%	75%	86%	85%	86%	85%

The evaluation results in table 1 show significant values for accuracy, precision, recall, and F1-score, particularly for the GRU-BiLSTM model with an 80:20 data split. When SMOTE is applied to achieve balanced classes, the model demonstrates more consistent performance across all metrics. High precision indicates that the model can avoid misclassifying certain emotions, which is crucial for maintaining accurate emotion interpretation in text. High recall shows that the model effectively detects various emotions, as the balanced classes reduce the risk of ignoring minority classes.

In comparison, in the test without SMOTE (class imbalance), the model tends to perform better on majority classes but is less optimal for minority classes. This is reflected in lower precision and recall for minority classes, indicating that the model struggles to detect emotions with lower frequencies in the imbalanced dataset. F1-score is also lower without SMOTE, showing that the model's performance is unbalanced and skewed towards majority classes. The use of SMOTE in this study ensures that the model is not only accurate for majority classes but also capable of effectively recognizing minority classes, resulting in a more comprehensive and fair performance overall.

4.2. Discussion

This research uses the SMOTE technique to overcome the problem of class integration in the tweet dataset used for sentiment analysis. Table 1 shows the results of testing deep learning models without using SMOTE, where the BiLSTM and hybrid GRU-BiLSTM models show lower accuracy and other performance metrics. In contrast, table 2 shows the same test results after applying SMOTE, which shows significant improvements in accuracy, precision, recall, and F1-score. Using SMOTE improves overall model performance because this technique balances the class distribution in the dataset by creating synthetic samples for minority classes. With more balanced data, the model can learn from more representative patterns from the entire data set, reducing bias towards the majority class. Increases the model's ability to identify and classify samples from minority classes more accurately. This performance improvement is consistent across different data splits (70:30, 80:20, and 90:10), demonstrating the effectiveness of SMOTE in various data split scenarios.

In addition, the number and type of layers used in the model also significantly influence performance. Deeper models with more layers, such as a combination of BiLSTM and GRU, can capture more complex features and the temporal context of the text. BiLSTM allows the model to capture context from two directions (forward and backward), improving understanding of the text. Meanwhile, GRU offers computational efficiency with a simpler structure than LSTM but can still handle long-term dependencies. Dropout layers are used as a regularization technique to prevent overfitting by randomly ignoring some neurons during training, which helps improve the model's generalization ability. Then, this research uses Glove as an embedding technique to convert tweet text into a meaningful vector representation. Glove produces vector representations that capture the semantic meaning of words well because it utilizes global statistical information from the entire text corpus. Allows the model to understand better the relationship between words and their context, which is very important in sentiment analysis tasks. By using Glove, the model has a richer and more meaningful representation of words for further processing. This research shows that the model can perform better sentiment analysis tasks by combining SMOTE, a GRU-BiLSTM hybrid architecture, and Glove embedding techniques. This results in more accurate and reliable predictions, and the accuracy is also better compared to previous research. Table 2 compares this research with previous research that used emotion classes.

Table 2. Comparison with Previous Research

Researcher	Model	Dataset	Accuracy
Mansy et al. [37]	BiLSTM	SemEval2018 task 1- Ec-Ar	54.00%
Setiawan & Andry [38]	Support Vector Regression	WASSA 2017 & SemEval2018	75.50%
Fahreza & Setiawan [39]	GRU	ISEAR	60.26%
Riza & Charibaldi [40]	LSTM	Twitter	73.14%
Ying et al. [41]	CNN	Twitter	77.59%
Bharti et al. [28]	CNN+BiGRU+SVM	ISEAR, WASSA, and Emotion-stimulus	80.11%
This Research	SMOTE+GRU-BiLSTM	Twitter	89.00%

Table 2 shows the test results of various deep learning models using SMOTE on various datasets for sentiment analysis. Research by Mansy et al. used the BiLSTM model on the SemEval2018 task 1 - Ec-Ar dataset and obtained an accuracy of 54%, indicating room for improvement in dealing with data complexity and class imbalance. Setiawan and Andry used Support Vector Regression on the WASSA 2017 and SemEval2018 datasets and achieved 75.5% accuracy, which is quite effective but can still be improved with other deep learning models. John et al. used GRU on the ISEAR dataset with an accuracy of 60.26%, indicating that the GRU model can be used for this task, but there is still room for performance improvement.

Research by Riza and Charibaldi used LSTM on the Twitter dataset and achieved 73.14% accuracy, showing better performance than several other models in this table. Ying et al. used CNN on the Twitter dataset. They achieved an accuracy of 77.59%, which shows that CNN can capture spatial features in text data, improving accuracy in sentiment analysis. Bharti et al., using a combination of CNN, BiGRU, and SVM on ISEAR, WASSA, and Emotion-stimulus datasets, achieved 80.11% accuracy. This model combination shows significant performance improvements by combining the strengths of multiple models.

This research uses a combination of SMOTE and the GRU-BiLSTM hybrid model for sentiment analysis on the Twitter dataset, achieving the highest accuracy of 89%. The use of SMOTE helps addresses the class imbalance in the dataset. At the same time, the combination of GRU and BiLSTM improves the model's ability to capture context and nuance in text, resulting in superior performance compared to other studies in this table. Thus, this study shows that the combination of SMOTE and the GRU-BiLSTM hybrid model provides the best performance in the sentiment analysis task on the Twitter dataset, with higher accuracy than other previous models.

4.3. Limitations of the Study

This study has several limitations that may impact the generalizability and applicability of the results. Firstly, the dataset used, primarily based on social media text, may not fully represent other types of text data, which can limit the model's adaptability to different contexts. Additionally, while SMOTE was applied to balance the classes, this approach may not perfectly capture the natural distribution of emotions in real-world data, which often tends to be highly imbalanced and complex. The model also demonstrated weaknesses in detecting certain emotions, such as 'surprise,' indicating the need for further optimization or alternative model architectures. The absence of temporal analysis means that changes in emotional context over time were not considered, which could be significant in dynamic situations like ongoing conflicts. For future research, the development of emotion detection models on social media can be expanded to applications in Decision Support Systems (DSS) and Internet of Things (IoT). These environments would require models capable of handling sensor and structured data in real-time, thereby supporting smarter and more responsive decision-making that adapts to contextual changes as they occur [42], [43].

5. Conclusion

This research shows that using the GRU-BiLSTM combination for sentiment analysis on tweets related to the Israeli-Palestinian conflict significantly improves model performance, especially when SMOTE balances data classes. Using Glove as an embedding technique ensures that the representation of words in the text has rich semantic meaning, which is crucial in understanding context and emotional nuances. The research results show that the GRU-BiLSTM hybrid model with SMOTE produces higher accuracy, precision, recall, and F1 scores than the model without SMOTE. Thus, this approach improves accuracy and provides fairer and more reliable predictions in sentiment analysis tasks on social media.

Future studies could consider utilizing advanced embedding techniques such as BERT or ELMo, which are capable of capturing word context more deeply through transformer architectures. These techniques enable the model to understand more complex contexts, especially in text containing sarcasm or irony, which are often challenging to classify accurately. Additionally, exploring model architectures such as Transformer or Attention-based LSTM could improve the model's sensitivity in capturing subtle emotions. Multi-task learning approaches could also be explored to combine emotion classification with other tasks, such as sentiment analysis, to enhance the model's generalization across various contexts.

6. Declarations

6.1. Author Contributions

Conceptualization: M.K.A., M., L.E., N.F., W.A., I.S., T.P.L., M.B.F., L., and A.K.S.; Methodology: I.S.; Software: M.K.A.; Validation: M.K.A., I.S., and M.B.F.; Formal Analysis: M.K.A., I.S., and M.B.F.; Investigation: M.K.A.; Resources: I.S.; Data Curation: I.S.; Writing Original Draft Preparation: M.K.A., I.S., and M.B.F.; Writing Review and Editing: I.S., M.K.A., and M.B.F.; Visualization: M.K.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. H. Elmasry, A. El Shamy, P. Manning, A. Mills, and P. J. Auter, "Al-Jazeera and Al-Arabiya framing of the Israel-Palestine conflict during war and calm periods," *Int Commun Gaz*, vol. 75, no. 8, pp. 750–768, Dec. 2013, doi: 10.1177/1748048513482545.

[2] S. Tezcan, "The Relations Between the Ottomans, Zionists and Palestinian Jews as Reflected in Israeli History Textbooks," *Belleten*, vol. 83, no. 298, pp. 1097–1130, 2019, doi: 10.37879/belleten.2019.1131.

[3] S. Adhim and Yuliati, "The Conflict of the Formation of the State of Israel in 1948-1973," *Asanka: Journal of Social Science and Education*, vol. 2, no. 1, pp. 61–70, 2021, doi: 10.21154/asanka.v2i1.2429.

[4] A. Khairunnisa, "Ideological Perlocutions of Palestinian and Israeli News on CNN Arabic Instagram (Pragmatic Cyber Study)," *International Journal of Islamic civilization*, vol. 6, no. 1, pp. 46–68, 2023, doi: 10.14421/skijic.v6i1.2844.

[5] R. W. Pratiwi, Y. Sari, and Y. Suyanto, "Attention-Based BiLSTM for Negation Handling in Sentimen Analysis," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 4, pp. 397–406, Oct. 2020, doi: 10.22146/ijccs.60733.

[6] Y. H. Hsieh and X. P. Zeng, "Sentiment Analysis: An ERNIE-BiLSTM Approach to Bullet Screen Comments," *Sensors*, vol. 22, no. 14, pp. 1–15, Jul. 2022, doi: 10.3390/s22145223. Link: https://www.mdpi.com/1424-8220/22/14/5223

[7] M. Chihab, M. Chiny, N. M. H. Boussatta, Y. Chihab, and M. Youssef Hadi, "BiLSTM and Multiple Linear Regression based Sentiment Analysis Model using Polarity and Subjectivity of a Text," I*JACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 436–442, 2022, doi: 10.14569/IJACSA.2022.0131052.

[8] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews," *J Sens*, vol. 2022, no. 1, pp. 1–12, 2022, doi: 10.1155/2022/7212366.

[9] X. Li, Y. Lei, and S. Ji, "BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Buzzwords," *Future Internet*, vol. 14, no. 11, pp. 1–15, Nov. 2022, doi: 10.3390/fi14110332.

[10] S. Wang, Y. Zhu, W. Gao, M. Cao, and M. Li, "Emotion-semantic-enhanced bidirectional LSTM with multi-head attention mechanism for microblog sentiment analysis," *Information (Switzerland)*, vol. 11, no. 5, pp. 1–15, Jun. 2020, doi: 10.3390/INFO11050280.

[11] A. Glenn, P. LaCasse, and B. Cox, "Emotion classification of Indonesian Tweets using Bidirectional LSTM," *Neural Comput Appl*, vol. 35, no. 13, pp. 9567–9578, May 2023, doi: 10.1007/s00521-022-08186-1.

[12] M. A. Alshamari, "Evaluating User Satisfaction Using Deep-Learning-Based Sentiment Analysis for Social Media Data in Saudi Arabia's Telecommunication Sector," *Computers*, vol. 12, no. 9, pp. 1-24, Aug. 2023, doi: 10.3390/computers12090170.

[13] B. Abimbola, E. de La Cal Marin, and Q. Tan, "Enhancing Legal Sentiment Analysis: A Convolutional Neural Network–Long Short-Term Memory Document-Level Model," *Mach Learn Knowl Extr*, vol. 6, no. 2, pp. 877–897, Apr. 2024, doi: 10.3390/make6020041.

[14] B. Gupta, P. Prakasam, and T. Velmurugan, "Integrated BERT embeddings, BiLSTM-BiGRU and 1-D CNN model for binary sentiment classification analysis of movie reviews," *Multimed Tools Appl*, vol. 81, no. 23, pp. 33067–33086, Sep. 2022, doi: 10.1007/s11042-022-13155-w.

[15] P. Durga, D. Godavarthi, S. Kant, and S. S. Basa, "Aspect-based drug review classification through a hybrid model with ant colony optimization using deep learning," *Discover Computing*, vol. 27, no. 19, pp. 1–27, Jul. 2024, doi: 10.1007/s10791-024-09441-w.

[16] M. K. Anam, M. B. Firdaus, F. Suandi, Lathifah, T. Nasution, and S. Fadly, "Performance Improvement of Machine Learning Algorithm Using Ensemble Method on Text Mining," in *ICFTSS 2024 - International Conference on Future Technologies for Smart Society*, Kuala Lumpur: Institute of Electrical and Electronics Engineers Inc., vol. 2024, no. Sept., pp. 90–95, Sep. 2024. doi: 10.1109/ICFTSS61109.2024.10691363.

[17] Z. Fitriyah and M. D. Kartikasari, "Text Classification of Twitter Opinion Related To Permendikbud 30/2021 Using Bidirectional Lstm," *BAREKENG*, vol. 17, no. 2, pp. 1113–1122, Jun. 2023, doi: 10.30598/barekengvol17iss2pp1113-1122.

[18] P. Sunagar and A. Kanavalli, "A Hybrid RNN based Deep Learning Approach for Text Classification," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 289-295, 2022, doi: 10.14569/IJACSA.2022.0130636.

[19] R. Refianti, A. B. Mutiara, and R. A. Putra, "A Lexicon-Based Long Short-Term Memory (LSTM) Model for Sentiment Analysis to Classify Halodoc Application Reviews on Google Playstore," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 146–157, Jan. 2024, doi: 10.47738/jads.v5i1.160.

[20] X. Li, Y. Zhang, J. Jin, F. Sun, N. Li, and S. Liang, "A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications," *PLoS One*, vol. 18, no. 3 March, pp. 1-20, Mar. 2023, doi: 10.1371/journal.pone.0282824.

[21] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1272–1285, Sep. 2024, doi: 10.47738/jads.v5i3.316.

[22] M. C. Mafunda, M. Schuld, K. Durrheim, and S. Mazibuko, "A word embedding trained on South African news data," *The African Journal of Information and Communication (AJIC)*, vol. 2022, no. 30, pp. 1-24, Dec. 2022, doi: 10.23962/ajic.i30.13906.

[23] A. Fellah, A. Zahaf, and A. Elçi, "Semantic Similarity Measure Using a Combination of Word2Vec and WordNet Models," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 455–464, Jun. 2024, doi: 10.52549/ijeei.v12i2.5114.

[24] D. Voskergian, R. Jayousi, and M. Yousef, "Enhanced TextNetTopics for Text Classification Using the G-S-M Approach with Filtered fastText-Based LDA Topics and RF-Based Topic Scoring: fasTNT," *Applied Sciences (Switzerland)*, vol. 14, no. 19, pp. 1–24, Oct. 2024, doi: 10.3390/app14198914.

[25] E. Cesario, C. Comito, and E. Zumpano, "A survey of the recent trends in deep learning for literature based discovery in the biomedical domain," *Neurocomputing*, vol. 568, no. 1, pp. 1–23, Feb. 2024, doi: 10.1016/j.neucom.2023.127079.

[26] H. Imaduddin, L. A. Kusumaningtias, and F. Y. A'la, "Application of LSTM and GloVe Word Embedding for Hate Speech Detection in Indonesian Twitter Data," *Ingenierie des Systemes d'Information*, vol. 28, no. 4, pp. 1107–1112, Aug. 2023, doi: 10.18280/isi.280430.

[27] M. K. Anam, I. Y. Pasa, K. D. kusuma Wardhani, L. Efrizoni, and M. B. Firdaus, "K-Means Clustering to Identity Twitter Build Operate Transfer (BOT) on Influential Accounts," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 14, no. 2, pp. 143–154, 2023, doi: 10.21512/comtech.v14i2.10620.

[28] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, A. Mahmoud., "Text-Based Emotion Recognition Using Deep Learning Approach," *Comput Intell Neurosci*, vol. 2022, no. 2645381, pp. 1–8, 2022, doi: 10.1155/2022/2645381.

[29] M. K. Anam, S. Defit, Haviluddin, L. Efrizoni, and M. B. Firdaus, "Early Stopping on CNN-LSTM Development to Improve Classification Performance," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1175–1188, 2024, doi: 10.47738/jads.v5i3.312.

[30] S. Choo and W. Kim, "A study on the evaluation of tokenizer performance in natural language processing," *Applied Artificial Intelligence*, vol. 37, no. 1, pp. 512-529, 2023, doi: 10.1080/08839514.2023.2175112.

[31] L. L. Van Fc, M. K. Anam, M. B. Firdaus, Y. Yunefri, and N. A. Rahmi, "Enhancing Machine Learning Model Performance in Addressing Class Imbalance," *COGITO Smart Journal*, vol. 10, no. 1, pp. 478–490, 2024.

[32] A. Angdresey and G. Saroinsong, "The Decision Tree Algorithm on Sentiment Analysis: Russia and Ukraine War," *Jurnal Sisfotenika*, vol. 13, no. 2, pp. 192–200, 2023, doi: 10.30700/jst.v13i2.1397.

[33] M. Ibrahim, S. Gauch, T. Gerth, and B. Cox, "WOVe: Incorporating Word Order in GloVe Word Embeddings," *International Journal on Engineering, Science and Technology*, vol. 4, no. 2, pp. 124–29, 2022, doi: 10.46328/ijonest.83.

[34] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artif Intell Rev*, vol. 56, no. 9, pp. 10345–10425, Sep. 2023, doi: 10.1007/s10462-023-10419-1.

[35] I. Salehin and D. K. Kang, "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain," *Electronics (Switzerland)*, vol. 12, no. 1, pp. 1–23, Jul. 2023, doi: 10.3390/electronics12143106.

[36] A. Zafar, M. Aamir, N. M. Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, S. Almotairi., "A Comparison of Pooling Methods for Convolutional Neural Networks," *Applied Sciences (Switzerland)*, vol. 12, no. 17, pp. 1–21, Sep. 2022, doi: 10.3390/app12178643.

[37] A. Mansy, S. Rady, and T. Gharib, "An Ensemble Deep Learning Approach for Emotion Detection in Arabic Tweets," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 980–990, 2022, doi: 10.14569/IJACSA.2022.01304112.

[38] R. C. Setiawan and A. Chowanda, "Emotion Intensity Value Prediction with Machine Learning Approach on Twitter," *CommIT Journal*, vol. 17, no. 2, pp. 235–243, 2023, doi: 10.21512/commit.v17i2.8503.

[39] Syafa Fahreza and E. B. Setiawan, "Sentiment Analysis on Social Media Using Word2Vec and Gated Recurrent Unit (GRU) with Genetic Algorithm Optimization," *International Journal on Information and Communication Technology (IJoICT)*, vol. 10, no. 1, pp. 62–77, Jun. 2024, doi: 10.21108/ijoict.v10i1.903.

[40] M. A. Riza and N. Charibaldi, "Emotion Detection in Twitter Social Media Using Long Short-Term Memory (LSTM) and Fast Text," *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 3, no. 1, pp. 15–26, May 2021, doi: 10.25139/ijair.v3i1.3827.

[41] O. J. Ying, M. M. A. Zabidi, N. Ramli, and U. U. Sheikh, "Sentiment analysis of informal malay tweets with deep learning," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 212–220, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp212-220.

[42] U. Rahmalisa, A. Febriani, and Y. Irawan, "Detector leakage gas LPG based on telegram notification using wemos D1 and MQ-6 sensor," *Journal of Robotics and Control (JRC)*, vol. 2, no. 4, pp. 287–290, Jul. 2021, doi: 10.18196/jrc.2493.

[43] Y. Irawan, "Decision Support System For Employee Bonus Determination With Web-Based Simple Additive Weighting (SAW) Method In PT. Mayatama Solusindo," *Journal of Applied Engineering and Technological Science*, vol. 2, no. 1, pp. 7–13, Nov. 2020, doi: 10.37385/jaets.v2i1.162.