All Sciences Proceedings

3rd International Conference on Innovative Academic Studies

September 26-28, 2023 : Konya, Turkey

ICIAS

All Sciences Proceedings
http://as-proceeding.com/

https://www.icias.net/

Social Media Sentiment Analysis Classification Leveraging Hybrid Deep Learning Methods

Ömer Ayberk ŞENCAN*, İsmail ATACAK [2]

[1] Gazi University, Technology Faculty, Dept. of Computer Engineering 06560 - Ankara, Türkiye
[2] Gazi University, Technology Faculty, Dept. of Computer Engineering 06560 - Ankara, Türkiye

*(oayberksencan@gazi.edu.tr)

Abstract – Social media platforms are one of the most popular platforms for users to express their opinions and sentiments concerning various products, services, and organizations. This study presents the models that can successfully analyze sentiment through text on social media platforms using hybrid deep learning methods. The proposed models were applied to the "U.S. Airline Dataset" obtained from the Kaggle platform. After the text cleaning phase, the BERT embeddings were implemented on the dataset. The resulting preprocessed data set was used in the training of three deep learning-based hybrid algorithms, namely, CNN-GRU, BiLSTM-GRU, and RNN-GRU. The experimental results revealed that the best result was achieved by the CNN-GRU model, with an accuracy of 0.79 and an F-Score of 0.79.

Keywords – Sentiment Analysis, Transfer Learning, Deep Learning, Social Media Analysis, Twitter

I. INTRODUCTION

Sentiment analysis is an essential part of artificial intelligence-based social media analysis. It has been utilized in different domains, such as politics, marketing, advertisement management, cybercrime detection, and many other areas of expertise [1].

The increase in internet usage has significantly increased the use of social media platforms, including Twitter [2]. A significant portion of their user base employs these platforms not solely for social networking and sharing personal experiences, but also as a means to express their perspectives and sentiments concerning various products, services, and organizations by engaging in discussions and making posts [3], [4]. Users employ Twitter as a platform to express their perspectives and provide feedback, encompassing both positive and negative sentiments, concerning the products and services they have acquired or utilized [5].

Analyzing consumer feedback is one of this domains. Smartphones and other devices have made social media platforms such as Twitter considerably more accessible for users with the advancement of the devices we use every day. This led the social media users to spend more time in this platform during the day which resulted in an exponential growth in the amount of data these users generated. Consequently, different approaches have been proposed in this time period to effectively process the raw social media data and obtain useful information using it. Machine learning-based approaches have had considerable success in this field quickly since it is nearly impossible to handle raw data using solely human effort. Even while most research questions can be solved using standard machine learning-based methods, processing raw text data, and extracting the user's sentiment from it is a difficult task. Therefore, using deep learning-based models, the researchers were able to overcome this obstacle.

Researchers of the existing studies in this domain mainly used either machine learning or deep learning-based methods [6].

Basiri et al. proposed a fusion model, using CNN, BiGRU, fastText, NBSVM and Distilbert with a Covid-19 dataset consisted of over 1 million tweets and obtained an accuracy of 0.858 and F-Score of 0.858 [7]. Salur and Aydin trained LSTM, GRU, BiLSTM, and CNN deep learning algorithms on tweet datasets consisting of 17,289 instances. As a result of this training, they achieved an accuracy metric of 0.821 with the CNN-BiLSTM model [5]. Rahman and Islam utilized a Tweet dataset comprising 12,000 instances to train machine learning-based models consisting of Decision Tree, SVM, Logistic Regression, and Stacking Classifier. The most successful result was observed with the Stacking Classifier, achieving an F-Score metric of 0.835 [8].

Furthermore, it has been observed in some studies in the literature that hybrid methods have also been employed [9]. In the study conducted by Zainuddin et al., a hybrid model was developed using three different tweet datasets. SentiwordNet lexicon, PCALSA/RP, Support Vector Machine, Naive Bayes, Random Forest, and Extreme Learning Machine were employed. The experimental results revealed that the most successful outcome was achieved using the Stanford Twitter Sentiment (STS) dataset, with an accuracy of 0.76 and an F-Score of 0.76 [10].

Table 1 provides an overview of some studies in the literature, presenting the datasets employed, the number of data samples, the models utilized, and the resulting performance metrics in these studies.

Table 1. Some studies in the literature with the performance metrics

Study	Dataset	Data Counts	Model	Performance Result
Basiri et al.	Covid-19	1056049	Proposed fusion model	Accuracy: 0.858
				F-Score: 0.858
Rahman and Islam	Tweets	12000	Stacking Classifier	F-Score:0.835
Salur and Aydin	Tweets	17289	CNN-BiLSTM	Accuracy: 0.821
Zainuddin et al.	Stanford Twitter Sentiment (STS)	353	Hybrid	Accuracy: 0.760
				F-Score: 0.760

II. Materials and Method

Kaggle is an online platform that hosts data science competitions and the datasets used in these events [7]. In this study, the raw data set is obtained from this platform. The dataset comprises 14640 tagged tweet data containing consumer feedback about six major US airlines. Figure 1 illustrates a few tweets taken from the raw data set.

airline_sentiment	text
neutral	@VirginAmerica What @dhepburn said.
positive	@VirginAmerica plus you've added commercials to the experience... tacky.
neutral	@VirginAmerica I didn't today... Must mean I need to take another trip!

Fig. 1 A few sample tweets.

The word cloud created using the raw data set is given in Figure 2.

Fig. 2 Wordcloud of raw dataset

In addition to the information provided above, the histogram used to examine the extent of the lengths of tweets in the data set is given in Figure 3.

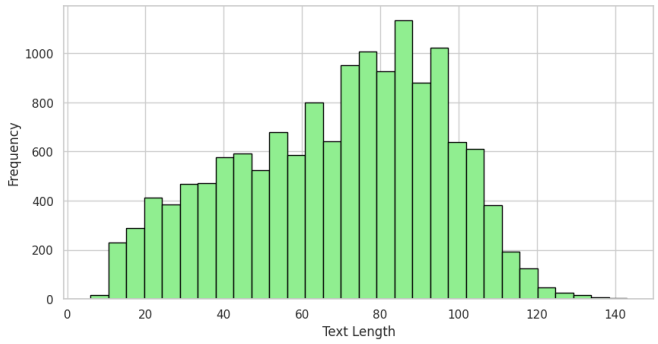

Fig. 3 Text Length Distribution

Figure 3 shows that the tweets in the data set mostly range between 60 and 100 words long. Furthermore, the following results were found when the most frequently occurring terms in the data set were examined in Table 2.

Table 2. Most common words and re-occurrence counts

Word	Number of occurrences
united	4144
flight	3895
usairways	3051
americanair	2957
southwestair	2453
jetblue	2361
get	1336
thanks	1072
cancelled	1056
service	956

After obtaining raw data, the data should go through a pre-processing stage before being fed into machine learning and deep learning-based models. All tweets in this context have been transformed into lowercase letters. Then, aspects in the tweets that did not alter the emotion included in the text content, such as special characters, numerals, and URLs, were dropped. After this procedure, stop words and punctuation marks were deleted from the text data, which are described in English as words that do not influence the content of the phrase, and the data was cleaned. The sentiment labels which constructed as text in the data set as positive, negative, and neutral are transformed as follows:

- Positive: 2,
- Neutral: 1,
- Negative: 0.

By this procedure, it is assured that all data is modified such that it may be fed into machine learning-based models by structuring it as given above. Figure 4 illustrates some instances of dataset produced as the result of preprocessing.

airline_sentiment	text
1	virginamerica dhepburn said
2	virginamerica plus youve added commercials experience tacky
1	virginamerica didnt today must mean need take another trip

Fig. 4 Example instances of dataset after preprocessing

When the data set reviewed after cleaning, it was discovered that the data counts on a class basis were as shown in Figure 5. Considering this condition will cause the data set to be unbalanced, lowering the performance of the machine learning and deep learning-based models that will be utilized, the data set needed to be balanced using synthetic data synthesis techniques. The SMOTE approach was employed to balance the data set in the above scenario.

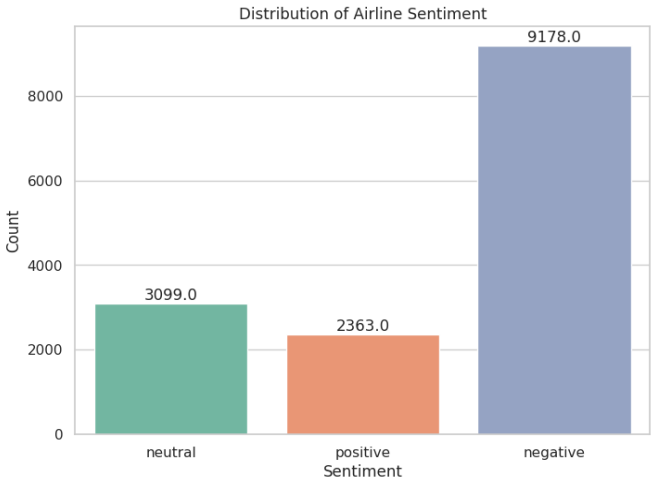

Fig. 5 Distribution of instances in the dataset by class after SMOTE

The preprocessed data set containing 27,561 values (9178 from each class) was generated following the SMOTE application. After obtaining the cleaned text data, all rows containing Null values were eliminated from the data set, and BERT embeddings were applied to the data set. After the application of BERT embeddings, the data set was transformed into a form that can be used in this study. Figure 6 shows the flowchart outlining the steps involved in the proposed approach in this study.

Following the preprocessing, the final data set was split into 75% training and 25% testing segments for utilization in deep learning-based models. The resulting preprocessed data set was used in the training of three deep-learning based hybrid algorithms, namely, RNN-GRU, CNN-GRU and BiLSTM-GRU. The experimental studies in the study were carried out using the Google Colab platform. The hyperparameters and settings used are given in Table 3.

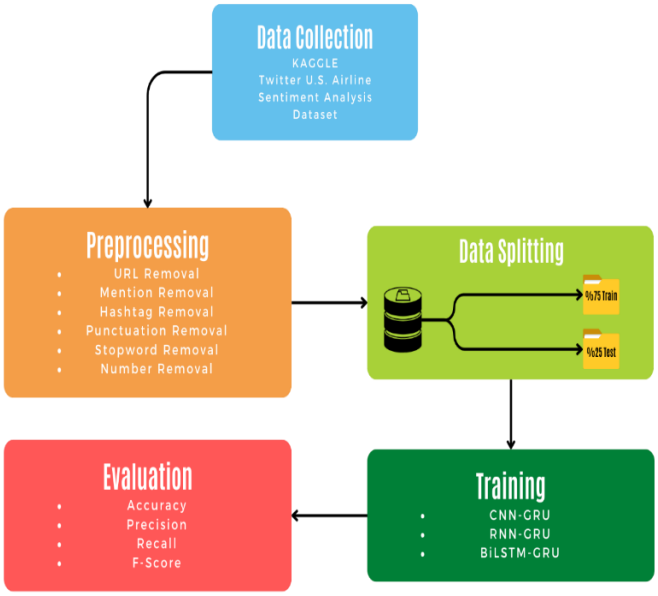

Fig. 6. Flowchart of the proposed approach

Table 3. Hyperparameters of proposed models

Model	Hyperparameters
RNN-GRU	Optimizer: Adam
	Loss Function: sparse_categorical_crossentropy
	Early Stopping Patience: 20 Epochs
	Early Stopping Restore Best Weights: True
	Epochs: 50
	batch_size: 64
	Activation Function: Softmax
CNN-GRU	Optimizer: Adam
	Loss Function: sparse_categorical_crossentropy
	Early Stopping Patience: 20 Epochs
	Early Stopping Restore Best Weights: True
	Epochs: 50
	batch_size: 64
	Activation Function: Softmax
BiLSTM-GRU	Optimizer: Adam
	Loss Function: sparse_categorical_crossentropy
	Early Stopping Patience: 20 Epochs
	Early Stopping Restore Best Weights: True
	Epochs: 50
	batch_size: 64
	Activation Function: Softmax

III. Results And Discussion

The results of the experimental study conducted utilizing the data set described in Section 2 are presented in this section. In order to determine the success rate of the proposed models, the experimental results are given in tables, as well as accuracy and loss graphs derived from the training phase.

Following the training process on the Google Colab platform with the hyperparameters listed in Table 2, the performance results obtained by the models employed in this study are shown in Table 4.

Table 4. Performance results of the models used in this study

Model	Accuracy	Precision	Recall	F-Score
CNN-GRU	0.79	0.79	0.79	0.79
RNN-GRU	0.72	0.72	0.72	0.71
BiLSTM-GRU	0.74	0.74	0.74	0.74

The experimental results reveal that the CNN-GRU model, which obtained the F-Score value of

0.79, achieved better performance results with the data set than both the RNN-GRU model, which obtained an F-Score value of 0.75 and the BiLSTM-GRU model with the F-Score value of 0.74. Figure 7 depicts the performance results of the proposed models per class.

According to Figure 7, the classification of negative instances in the data set was performed more successfully than the other two classes in all models with the average F-Score value of 0.81. Negative instances were followed by Positive, with the average F-Score value of 0.75, and Neutral examples, with the average F-Score value of 0.68. In this context, the most successful result was obtained by the CNN-GRU model, which classified negative data with the F-Score value of 0.86. Figure 8 illustrates the F-Score values of each model for each class.

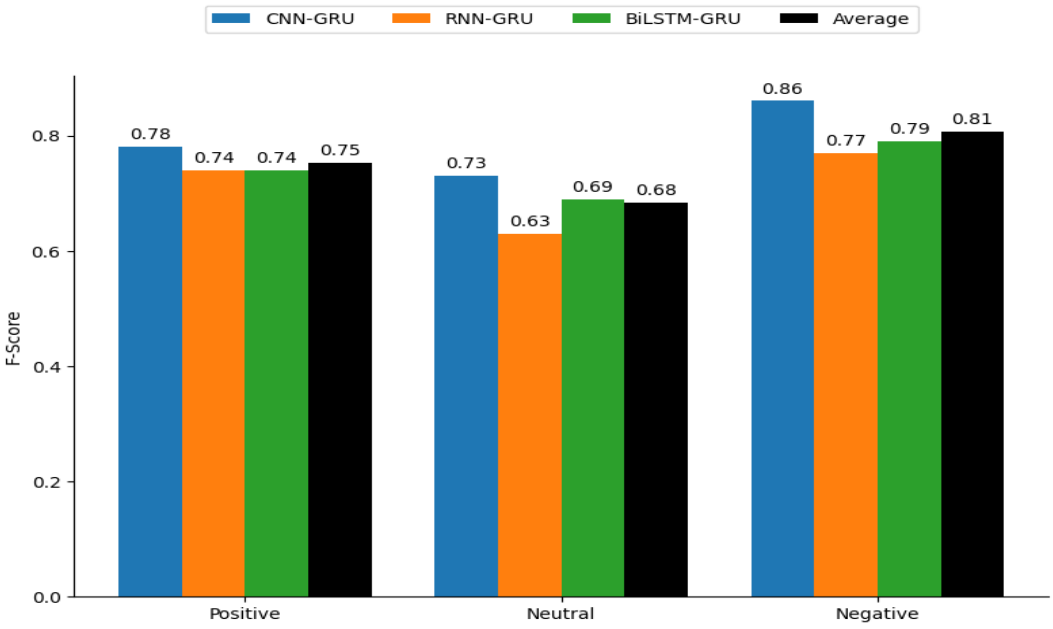

Fig. 7. Model performance by class

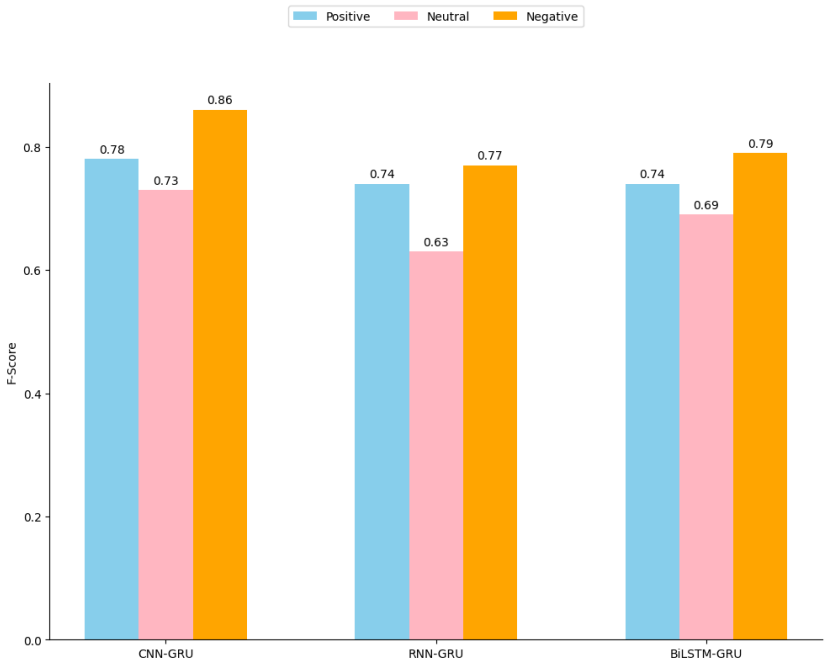

Fig. 8. Class performance by model

According to the data in Figure 8, the BiLSTM-GRU model achieved the most successful results across all classes. This model achieved the F-Score value of 0.86 for the Positive class, 0.77 for the Neutral class, and 0.79 for the Negative class. In addition to the information provided above, accuracy and loss graphs for training and validation acquired by the proposed models during training phase are provided below in Figure 9, Figure 10 and Figure 11.

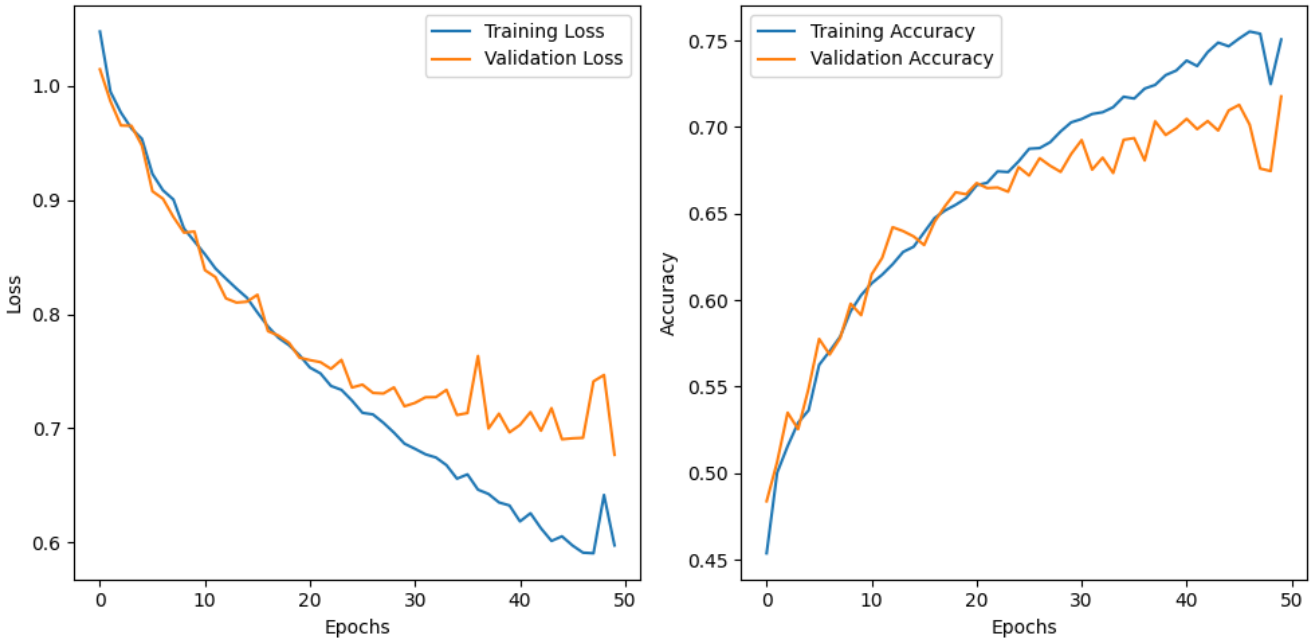

Fig. 9. Training graph of BiLSTM-GRU model

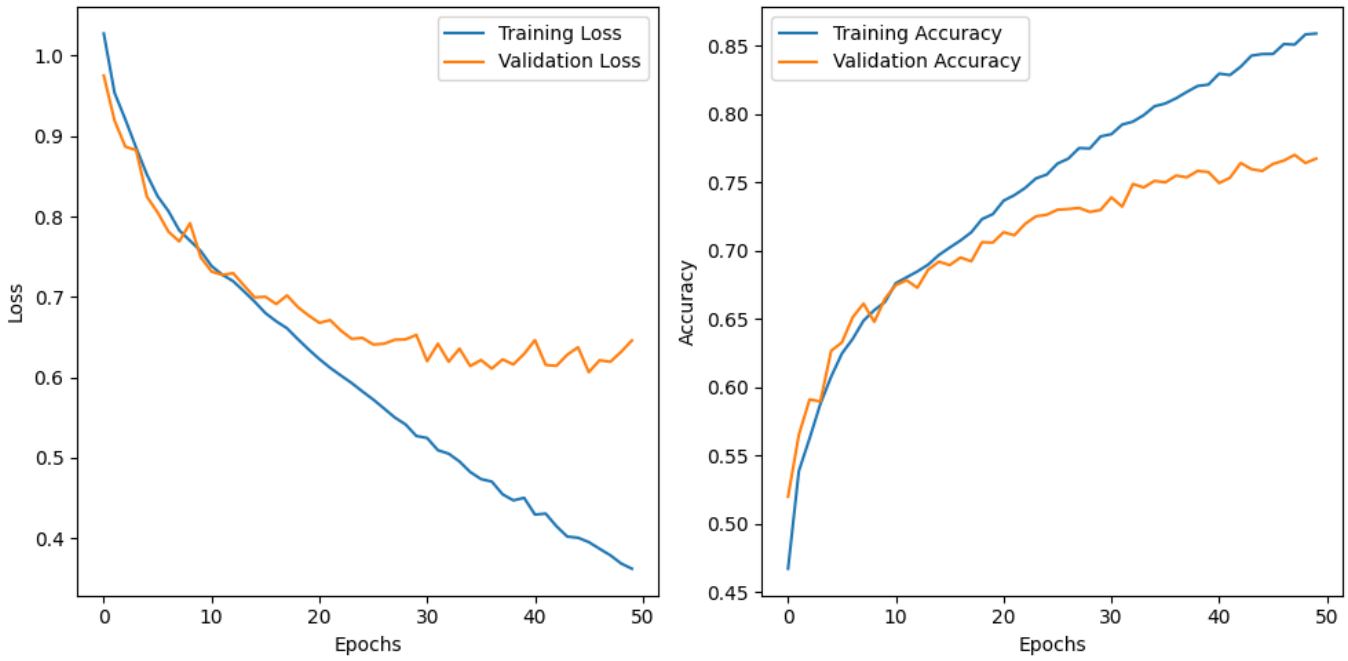

Fig. 10. Training graph of CNN-GRU model

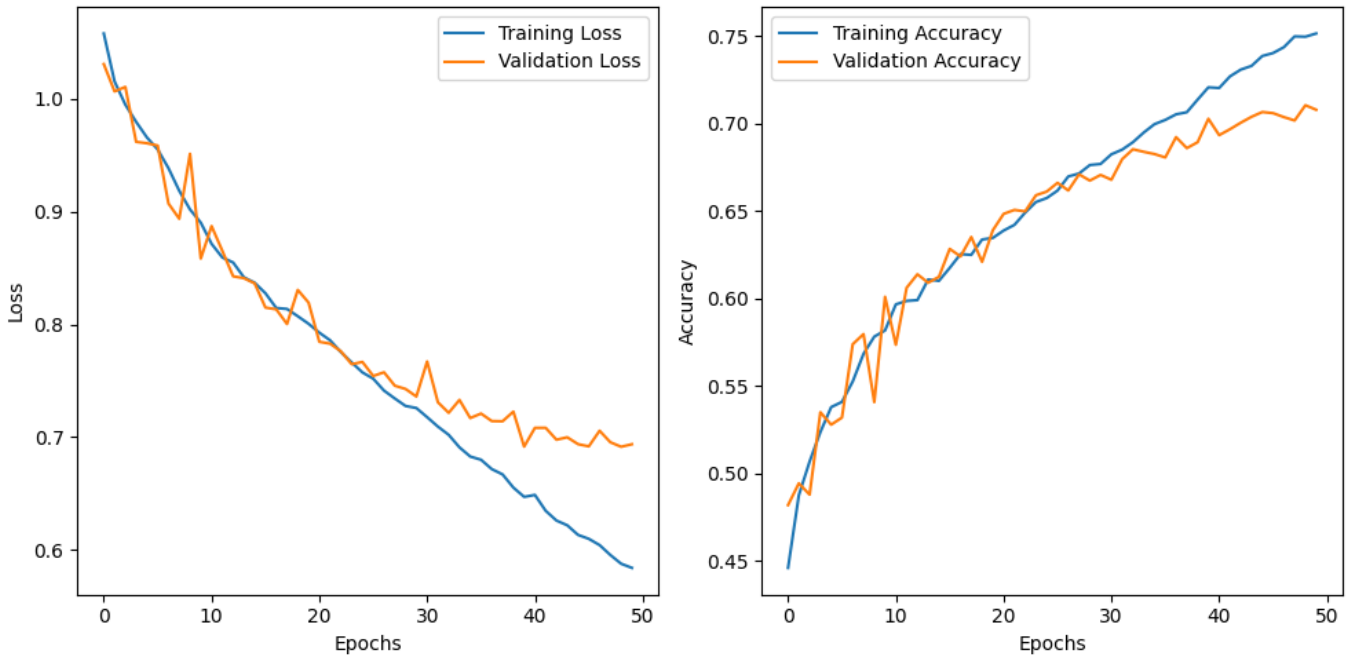

Fig. 11. Training graph of RNN-GRU model

Figure 9, Figure 10 and Figure 11 show the accuracy and loss graphs of the BiLSTM-GRU, CNN-GRU and RNN-GRU models in the training phase, respectively. Loss values in these graphs are automatically calculated by the system using the loss function.

IV. CONCLUSION

This study presents the hybrid deep learning based models that perform sentiment analysis on tweet text using the "U.S. Airline Dataset" data set obtained from the Kaggle platform. Pre-trained BERT embeddings were utilized in the development of the proposed models. Precision, Recall, and F-Score metrics, frequently employed in the literature, were used in performance evaluations of the models. Furthermore, these evaluation results were confirmed using graphs that incorporate accuracy and loss values obtained for the training and validation phases. In this context, when the results obtained by the proposed models were examined, it was seen that they could successfully analyze sentiments through text on social media platforms. The results also serve as evidence of the contribution of this study to the existing literature.

REFERENCES

[1] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst Appl*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/j.eswa.2023.119862.

[2] A. Perti, M. C. Trivedi, and A. Sinha, "Development of intelligent model for twitter sentiment analysis," *Mater Today Proc*, vol. 33, pp. 4515–4519, 2020, doi: 10.1016/j.matpr.2020.08.004.

[3] M. M. Agüero-Torales, M. J. Cobo, E. Herrera-Viedma, and A. G. López-Herrera, "A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor," *Procedia Comput Sci*, vol. 162, pp. 392–399, Jan. 2019, doi: 10.1016/J.PROCS.2019.12.002.

[4] N. AL-Bakri, J. Yonan, A. S.-B. S. Journal, and undefined 2022, "Tourism companies assessment via social media using sentiment analysis," *iasj.netNF AL-Bakri, JF Yonan, AT SadiqBaghdad Science Journal, 2022•iasj.net*, 2021, doi: 10.21123/bsj.2022.19.2.0422.

[5] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/ACCESS.2020.2982538.

[6] S. Khatoon and Lamis Abu Romman, “Domain Independent Automatic Labeling system for Large-scale Social Data using Lexicon and Web-based Augmentation,” *Information Technology And Control*, vol. 49, no. 1, pp. 36–54, Mar. 2020, doi: 10.5755/j01.itc.49.1.23769.

[7] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, “A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets,” *Knowl Based Syst*, vol. 228, p. 107242, Sep. 2021, doi: 10.1016/j.knosys.2021.107242.

[8] Md. M. Rahman and M. N. Islam, “Exploring the Performance of Ensemble Machine Learning Classifiers for Sentiment Analysis of COVID-19 Tweets,” 2022, pp. 383–396. doi: 10.1007/978-981-16-5157-1_30.

[9] Q. A. Xu, V. Chang, and C. Jayne, “A systematic review of social media-based sentiment analysis: Emerging trends and challenges,” *Decision Analytics Journal*, vol. 3, p. 100073, Jun. 2022, doi: 10.1016/J.DAJOUR.2022.100073.

[10] N. Zainuddin, A. Selamat, and R. Ibrahim, “Hybrid sentiment classification on twitter aspect-based sentiment analysis,” *Applied Intelligence*, Dec. 2017, doi: 10.1007/s10489-017-1098-6.

[11] C. S. Bojer and J. P. Meldgaard, “Kaggle forecasting competitions: An overlooked learning opportunity,” *Int J Forecast*, vol. 37, no. 2, pp. 587–603, Apr. 2021, doi: 10.1016/J.IJFORECAST.2020.07.007.