Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Leveraging Deep Learning for Sentiment Analysis of Airline Customer Reviews and Feedback

P. Naveen Sundar Kumar[1], K. Manasa[2], M. Sree Neeha[2], B. Likitha Reddy[2], B. Sainath[2]

[1]Assistant Professor, Dept. of Computer Science and Engineering, RGM College of Engineering and Technology, Nandyal (Dist), AP, India.
[2]Students, Dept. of Computer Science and Engineering, RGM College of Engineering and Technology, Nandyal (Dist), AP, India.
Email: nvnp03@gmail.com,

Sentiment analysis can identify the emotional tone of textual data, including whether it is favorable, negative, or neutral. One use of natural language processing (NLP) is sentiment analysis. In order to improve corporate strategy and services, client feedback is essential. Analyzing client feedback on social media is crucial for providing businesses with more accurate results. This study suggests a deep learning system for sentiment analysis that makes use of RNN, LSTM, and GRU models. Data preprocessing, feature extraction, model training, and evaluation are some of the several processes that are involved. The system guarantees more accurate and suitable classification by examining each model's advantages. The suggested method yields notable performance gains, making it robust and flexible solution for classifying and analyzing sentiments in large-scale airline customer reviews.

Keywords: Sentiment Analysis, Deep Learning, Airline Customer Reviews, Natural Language Processing.

1. Introduction

IRLINES always operate in a highly competitive market where every experience matters to improve their services. Customer feedback is most influential in the complete aviation industry, as it gives crucial insights so as to improve customer satisfaction, expectations and also provides more scope to identify different areas of improvement. Reviewing passenger feedback helps airline to improvise their services, upgrade operational efficiency, and maintain brand loyalty. Web based social networks specially twitter emerged as leading medium for customers to share up to date reviews, their opinions and particularly our target is to capture emotions regarding their travel experiences. These reviews reflect both positive, negative and neutral sentiments, making Twitter a source of data for airlines to analyze and act accordingly.

These twitter feedbacks provides airlines a unique facility to gauge customer satisfaction and identify specific areas of trouble such as delays in-flight services, ticketing issues or overall comfort by evaluating these review airlines can enhance service gaps, observe passenger experiences and take active steps to solve customer issues in addition to this the twitter data allows us to keep a track on sentiment trends over time helping them to adapt their changes to meet the customer new expectations and remain competent in the market.

An application of natural language processing (NLP), sen- timent analysis is the act of analyzing textual data to ascertain the emotional tone—whether it be neutral, negative, or posi- tive. Sentiment analysis identifies various patterns of opinions and emotions, enabling firms to extract valuable insights from unstructured feedback. With the use of machine learning, deep learning, and natural language processing techniques, this has developed into a crucial tool for analyzing consumer per- ception, empowering companies—including airlines—to make data-driven decisions. Even in the face of ironic or humorous structured language, sentiment analysis can effectively inter- pret client attitudes.

Traditionally this is completely relied on rule-based methods that use some of the predefined dictionaries and sentiment lexicons for text classification. These simple and fast methods often strive to handle some linguistic expressions in opposition to this modern approaches that manipulates machine learning and deep learning techniques like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Transform- ers like BERT which are capable of understanding complex sentence structures and capturing long-term dependencies. These methods pose a significant improvement in accuracy by learning contextual nuances from huge datasets.

By implementing sentiment analysis airlines can improve their services in various ways including noticing customer complaints more effectively, improving flight experiences and improving staff support analyzing passenger sentiments helps airlines to discover hidden issues, predict customer dissatis- faction and prioritize actions that improve overall passenger experience. Sentiment analysis also enables airlines to monitor the impact new campaigns assuring continuous improvement in their services.

In this paper, we suggest a deep learning-based method for sentiment analysis of Twitter data that includes sophisticated models including RNN, LSTMs, and GRUs. These methods are highly accurate in capturing the sequential and context information of text data, enabling correct classification of customer sentiments. By using this approach we aim to improve the aviation industry with more insights that helps airlines enhance their services to increase customer satisfaction and maintain competitive spirit.

2. Literature Review

Yasmin Yashodha (2012) undertook the business level, corporate level, and competitive-level strategies of AirAsia Berhad. In evaluating its success levels in penetrating unserved ASEAN markets and its capacity to maintain its competitive edge, this research utilised

PESTEL and Porter’s Industry Analysis frameworks, compiled both primary and secondary data, interviews inclusive.

In fact, Yun Wan and Qigang Gao (2015) have done a study on sentiment analysis within the airline service industry using twitter data. The study used ensemble classification approach by integrating Naiıve Bayes, SVM, Bayesian Network and Decision Trees algorithms. The ensemble approach with a gen- eralization accuracy of 75% implemented in 10-fold evaluation based on 12,864 tweets outperformed individual classifiers.

Liau, Bee Yee & Tan, Pei Pei (2016) examined the attitudes of customers towards LCCs in Malaysia by analyzing 10,895 tweets sourced in a sentiment analysis. Typically applying text mining and clustering models, including K-Means, they determined primary themes, like customer services, tickets offers and flight disruptions. The findings revealed that there was in general a sentiment polarity which was positive and this was spread across a number of different algorithms.[1]

Several authors, including Guoning Hu et al. (2017) studied sentiments of 330,000,000 tweets which included 19,000,000 users and 62 industries. Therefore, the study showed how the negative perception was most expressed in the service industry and how perception differences existed even between the brands and the industries. In the study, particular attention was paid to the possibility of using data collected through SNS in the assessment of customer attitudes and market conditions.

T. Hemakala and S. Santhoshkumar came up with the sentiment analysis framework based on tweets from six airlines operating in India in 2018. The preprocessing involved in this work are tokenization and text cleaning, as used in classification from Decision Tree, Random Forest and SVM. The outcomes of this research proved the capability of the mentioned methodologies for extending multi-class sentiment analysis to the airline sector.

There were Wajdi Aljedaani and Furqan Rustam who used the sentiment analysis on the corpus of 14,640 tweets from six US airlines. By applying TF-IDF and LSTM approach, the authors secured an accuracy of 92% as well as 97% thus asserting the applicability of deep learning in classifying tweets.[2]

Using both Adaboost and Random Forest, E. Prabhakar et al. [3] conducted sentiment analysis on the tweets of the top ten US airlines. Data preparation was the first phase in the experiment, followed by data split into the 75% and 25% groups division for testing and training. Adaboost gained the best values of precision (78%), recall and F-score (65%) comparative to the other classifiers investigated in this research work, followed by Random Forest which has the precision of 71%.

For sentiment analysis, Ankit and Nabizath Saleena [4] used an ensemble classification consisting of SVM, Random Forest, and Naive Bayes as basic classifiers. In this work, the Bag-of- Words technique was used to carry out the feature extraction step. It was discovered that the ensemble classifier produced sentiment ratings based on the likelihood of a positive or negative categorization and was more accurate.

Akshada Shitole and Archana Vaidya [5] tested four al- gorithms namely Boosting, Support Vector Machine (SVM), Decision Tree, and Logistic Regression to identify 14,640 tweets related to airlines. Part of the preprocessing was TF- IDF feature extraction. The overall result revealed that SVM have the highest accuracy of 90.7% and Decision Tree has the lowest of 79%.

Raihen and Akter [6] explored sentiment analysis for passenger feedback on six major U.S. airlines using seven machine learning classifiers: LDA, QDA, KNN, DT, RF, GBC, AdaBoost. The dataset contained 14,640 Twitter posts classified as Positive 2,363, Negative 9,178 and Neutral 3,099. Data preprocessing was done by converting the data to lower case, filtering out special characters and URLs and stopwords and splitting the data into tokens and finding the TF-IDF of the data. After analyzing the outcomes, it was assumed that Random Forest with 10 fold cross validation gives the highest accuracy of 90.13% compared to others. Therefore, it may be considered the most efficient classifier for this problem. The study shows that feature combination methods can notably improve the results of sentiment analysis, such as the RF method. From this analysis, airlines are able to understand areas of service which needs to be strengthened and customer satisfaction to be enhanced.

Using Yelp data, Patel et al. [7] compared the results of performing airline customer reviews. Sentiment analysis using Google's BERT model together with algorithms like Naive Bayes, SVM, DT and Random Forest. This dataset containing the text reviews has been obtained from Kaggle where text is classified as positive, negative or neutral. Steps of preprocess- ing included tokenization, punctuations removal and stop word removal. The entries were preprocessed to extract features by using the TF-IDF technique and classification schemes were posed on the data. The research analysis suggested that the BERT model was more effective in all parameters than the other models it provided an 83% accuracy compared to the 77% accuracy of the most effective traditional algorithm, Random Forest. One of the reasons for that was BERT's bidirectional learning and dynamic word embeddings. The study pointed to more experimental advanced BERT for further explorations as the study focused on the ability of BERT to perform the complex sentiment analysis tasks.

Sharma et al. [8] investigated sentiment analysis of movie reviews using three machine learning classifiers: Naive Bayes, Logistic Regression and Support Vector Machine (SVM). The study employed a dataset of 50,000 IMDb reviews labeled as positive or negative. Preprocessing steps included stopword removal, stemming and feature extraction using a Bag-of- Words (BoW) approach. The performance of the classifiers was evaluated using metrics such as accuracy, precision, recall and F1-score. SVM outperformed the other classifiers achieving the highest accuracy of 73%. This study highlighted the challenges of sentiment analysis in capturing subtle ex- pressions of sentiment and emphasized the utility of SVM for classifying movie reviews effectively.

For sentiment analysis Loh et al. presented a novel hybrid model of MPNet, GRU and BiGRU. Masked and permuted pre-training to improve contextual awareness is incorporated in the MPNet whose foundation is the transformer-based pre- trained language model. There is an efficient capturing of long-term dependencies when using GRU and bidirectional context is taken care of in BiGRU. Performance on IMDb, Twitter US Airline Sentiment and Sentiment140 dataset was 94.71%, 86.27% and 88.17% respectively. Some of the pre-treatment processes that were performed were tokenization, removal of stop words, normalization. Experimental outcomes provided evidence of the benefits of the model over the current approaches in addressing multiple and heterogeneous sentiments.[9]

The authors Steinke et al.[10] discussed the sentiment classification of movie reviews detected using various data mining techniques with the help of Decision Trees, Random Forests, and SVMs. To this end, the study employed Stanford Large Movie Review Dataset comprising 50,000 reviews alongside another dataset of the IMDb review data of 2019 and 2020. There was text preprocessing which included tokenization, removal of stopwords and stemmers plus the identification of feature using TF-IDF techniques. SVM has the highest percentage accuracy of 86.18% from the train dataset while Random forest has a percentage accuracy of 85.27%. The study also focused on how the COVID-19 pandemic might affect the sentiment there was a shift toward negative in 2020 though statistically the changes cannot be linked to the pandemic.

3. Proposed Methodology

A. Dataset Information

The dataset utilized in this study was taken from Kaggle and includes 14640 rows with 15 columns, each of which has a variety of features similar to tweets about airlines. This dataset is appropriate for sentiment analysis tasks since it offers a comprehensive picture of user thoughts, interac- tions, and related metadata. The tweets in this dataset are categorized using a number of criteria. The airline sentiment column specifically classifies tweets into three categories: favorable, neutral, and negative. In addition to the sentiment classification, the airline sentiment confidence feature offers a confidence score. The cause for discontent is specified in the negativereason column that is provided for a collection of negative tweets, and the confidence level in that classification is shown by the negativereason confidence column.

The text contains the actual content of given tweets that serves as the main input for sentiment analysis. In addition to this the dataset includes metadata such as the name of the air- line (airline), the user who tweeted (name), and the timestamp when the tweet was created (tweet created). Location-specific details such as tweet coord and tweet location, provide geographical context though these attributes have missing values for several records.

Two key binary attributes are airline sentiment gold and negativereason gold that provides gold standard labels for a little portion of the dataset useful for making outstanding models. The dataset also has various numeric features such as retweet count, representing the number of times a tweet was shared.

This dataset has challenges with real-world data such as handling missing values in features like negativereason, tweet coord and tweet location. The combination of categorical and numerical values when combined with textual data enables a accurate analysis but also gives best preprocessing and Deep learning techniques for accurate sentiment classification.

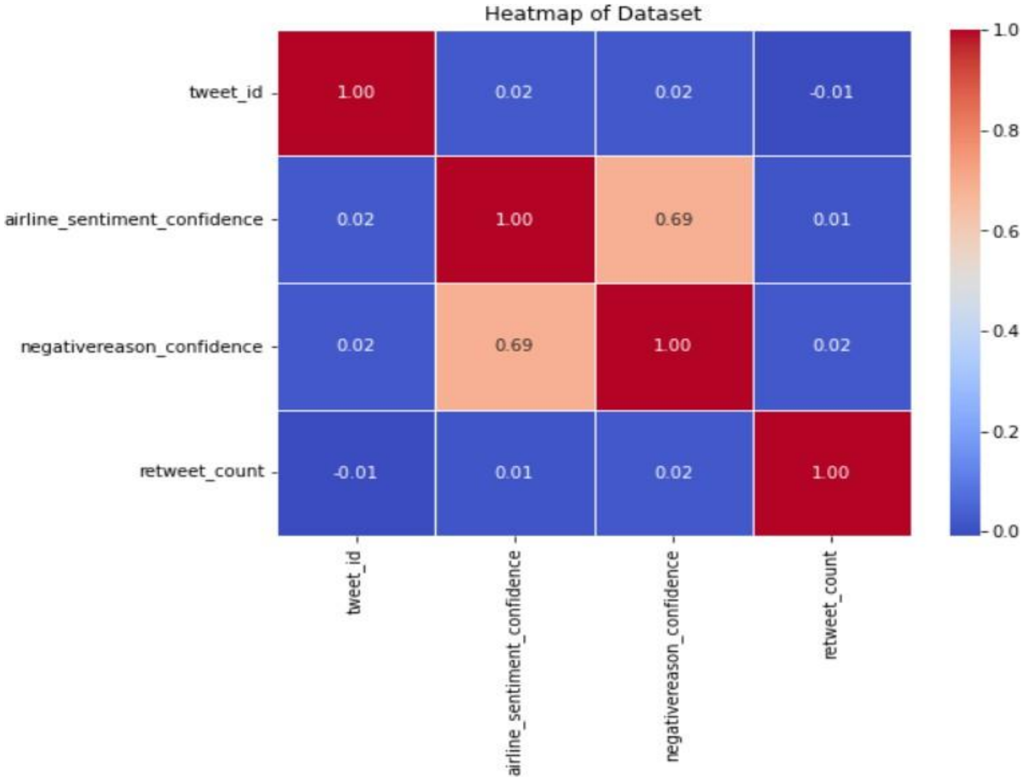

Fig. 1. Heat Map of dataset.

B. Data Preprocessing

This is the most essential step in preparing the dataset for sentiment analysis. The raw data has various issues like missing values, irrelevant features and also unstructured data that requires systematic cleaning and needs some transforma- tion.At first all the unrelated columns which includes tweet id, airline sentiment gold, negativereason gold and tweet coord are removed as they will not contribute to the sentiment classification task. Missing values in useful columns such as s negativereason were handled by replacing them with empty strings to maintain data integrity.

The text data in text column will undergo several cleaning steps. These steps include the removal of usernames, URLs, emojis and special characters, which are considered as extra data to the sentiment analysis. Short forms are expanded to their full forms to use a standard language and stop words were removed to use only meaningful words. The text was further nurtured to reduce words to their baseforms and also consistency is measured.

To change the cleaned text into a trainable format, the text was tokenized and converted to integer sequences. These sequences were then trimmed and padded to 100 tokens so that uniformity is maintained across the dataset. Sentiment labels were encoded into numeric values, where 0 represented negative, 1 represented neutral and 2 represented positive sentiments. This preprocessing dataset ensured that the dataset was clean, consistent and ready for input into the Deep learning models.

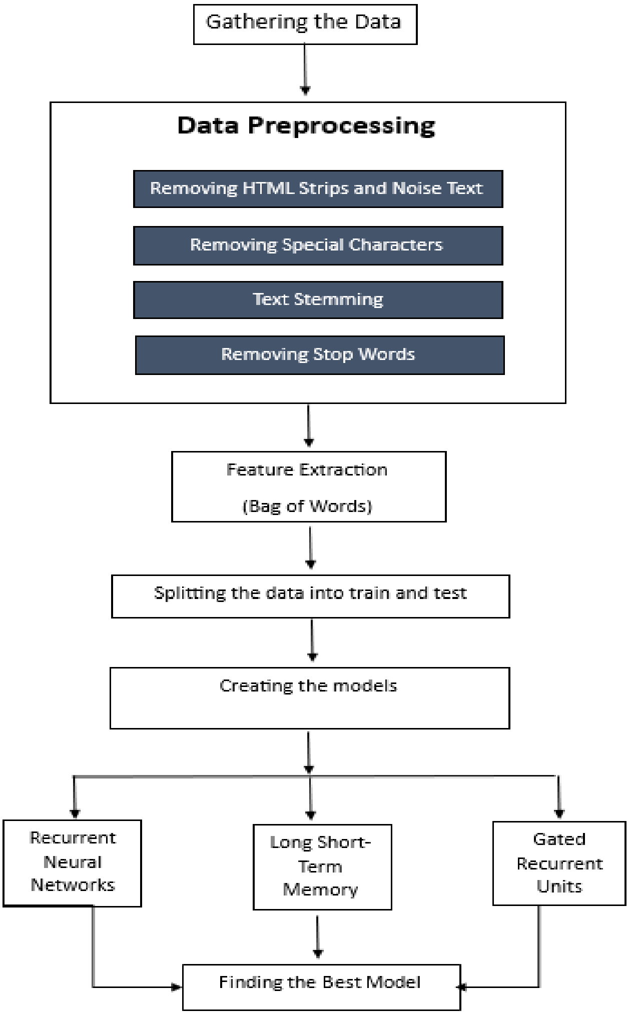

Fig. 2. Process Flow.

A. Model Architecture

The three deep learning architectures used to conduct sentiment analysis were Recurrent Neural Networks(RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). All the models focused on data that came in a sequence; this made it possible for them to conduct effective sentiment Classification.

The RNN model is the foundational architecture having an embedding layer to convert words into dense vectors, later it is followed by a bidirectional RNN layer that processed the sequence in both forward and backward directions. A fully connected dense layer and a softmax activation function were used to show the output of the probabilities for each sentiment class. The RNN's gating mechanism is mathematically defined as follows:

$$h_t = \sigma(W_h \cdot [h_{t-1}, x_t] + b_h) \qquad (1)$$

where h_t represents the hidden state at time t, x_t is the input at time t and σ is an activation function such as tanh or ReLU. The LSTM model is the moderate version of the RNN by incorporating memory cells that capture long-term dependencies. The LSTM's gating mechanism is mathematically defined

as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Here, f_t, i_t, o_t and C_t represent the forget gate, input gate, output gate and cell state respectively, while h_t is the hidden state.

The GRU model is the next version of the LSTM structure by combining the forget and input gates into a single update gate, represented as:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (7)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (8)$$

$$\tilde{h}t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \quad (9)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}t \quad (10)$$

Here, z_t and r_t are the update and reset gates, and $\tilde{h}t$ is the candidate hidden state.

B. Training and Validation

The training and evaluation process is completely designed to maximize the performance of the models and check their ability to classify sentiments accurately. The dataset splits into two sections: a training set and a testing set. The training set makes up 80% of the data, while the testing set accounts for the remaining 20%. During this training, the models are made to learn to map input sequences to sentiment labels by minimizing the error using the sparse categorical cross entropy loss function. This process was carried out using the rmsprop optimizer with a learning rate of 0.001, ensuring efficient output.

These models are trained over multiple datasets and epochs, with each epoch involving iterative updates to the model weights based on the error gradient. Batch processing was used to handle the large dataset effectively. To prevent overfitting and so as to improve generalization techniques such as dropout regularization were applied to the neural network layers.

After training the models are examined on testing dataset, which was kept not seen during training to provide an unbiased measure of their performance. The performance of each model RNN, LSTM and GRU was compared, and their respective strengths and disadvantages were examined to identify the most suitable architecture for sentiment analysis in this con- text.

E. Evaluation Metrics

Key evaluation metrics included accuracy, precision, recall and F1-score. These metrics provided a comprehensive understanding of the models ability to classify sentiments, balancing both correctness and robustness across all sentiment classes.

Formulas

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (11)$$

$$Precision = \frac{TP}{TP + FP} \qquad (12)$$

$$Recall = \frac{TP}{TP + FN} \qquad (13)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (14)$$

$$ERR = \frac{FP + FN}{P + N} \qquad (15)$$

4. Results

To evaluate the performance of the deep learning models (Recurrent Neural Network, Long Short Term Memory, Gated Recurrent Unit), following metrics were used: Accuracy, Precision, Recall, F1-score. For each cluster Gated Recurrent unit has given better results when compared with Recurrent Neural Network and Long Short Term Memory.

TABLE I. COMPARISON OF MODELS WITH ACCURACY, F1 SCORE, PRECISION, AND RECALL.

Sl.No	Model	Accuracy	F1 Score	Precision	Recall
1	RNN	0.929645	0.929442	0.937640	0.929645
2	LSTM	0.965000	0.965119	0.965423	0.965000
3	GRU	0.970000	0.969270	0.971892	0.970000

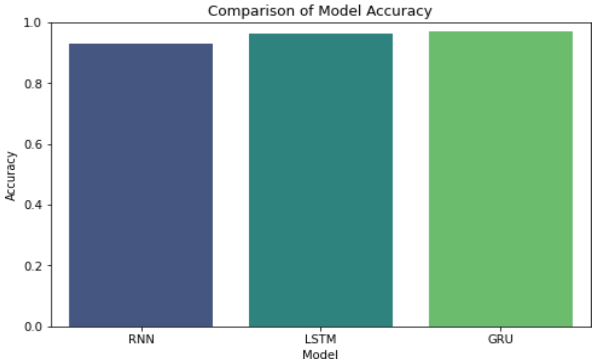

Fig. 3. Accuracy Graph.

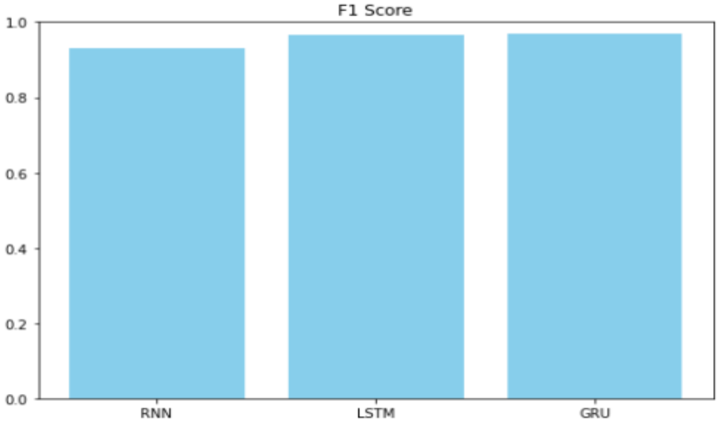

Fig. 4. F1 Score Graph.

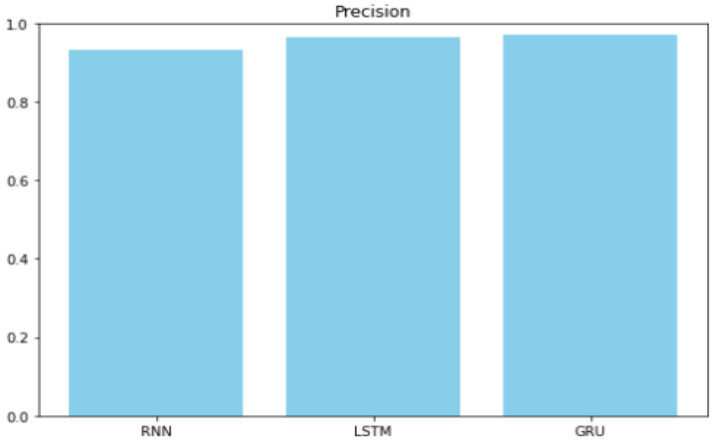

Fig. 5. Precision Graph.

In the process of sentiment analysis, the dataset that was collected earlier was preprocessed and missing values were handled. Then, a model was trained using techniques Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), and their evaluation metrics were compared to identify the best technique for sentiment analysis.

From the table and bargraphs we can analyse that Gated Recurrent Unit has given better performance than other models.

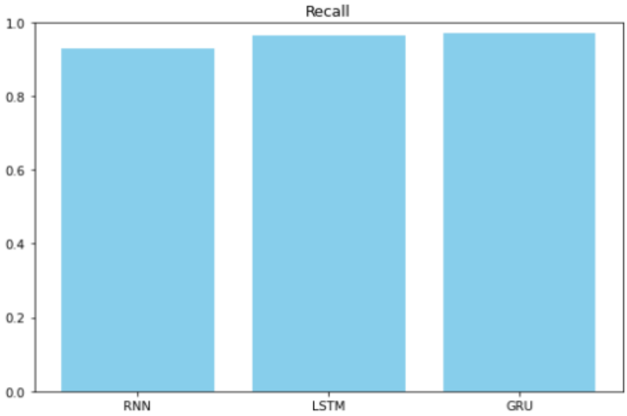

Fig. 6. Recall Graph.

5. CONCLUSION

In this paper, the study has established the potential of Deep Learning techniques in detecting airline customer feedback and reviews sentiment.

The models developed in deep learning had given better accuracy than traditional methods. Through data analysis, data handling, data preprocessing, model selection and training model automated the work of sentiment analysis of tweets given by airline customers. Out of all the models, Gated Recurrent Unit showed the highest accuracy, i.e., 97%.

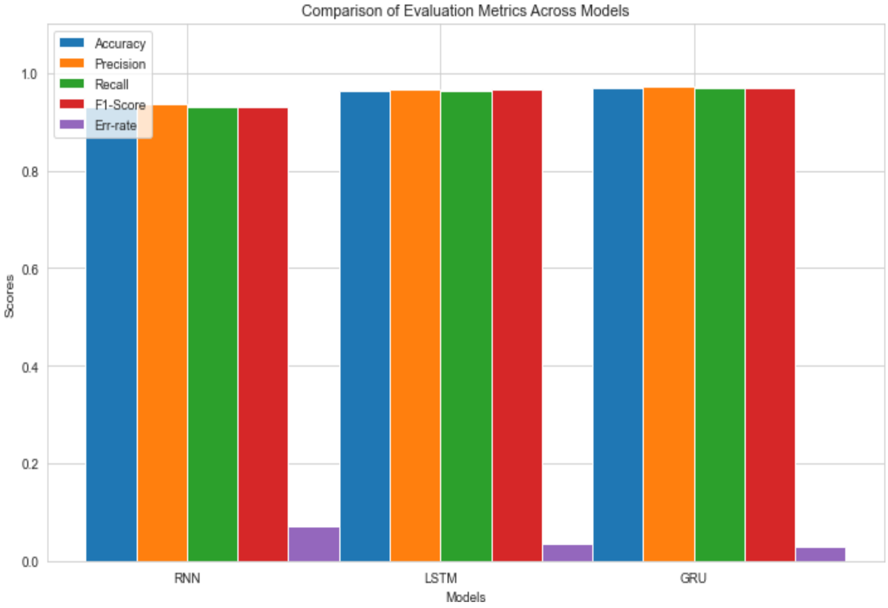

Fig. 7. Comparison graph between all the models.

A performance comparison of ML and DL models was conducted with a difference noted in the accuracy level achieved. Among the tested ML models, BERT provides the highest accuracy of 83%, which is higher than both Logistic Regression, KNeighbors, and Support Vector with accuracy around 65% – 67%. Extensions to these are the Random Forest and the Adaboost techniques that give better results of 77% and 72% respectively, based on enhanced model prediction. It can, therefore, be observed that Deep Learning models such as RNN, LSTM, and GRU have much better accuracy of 93%, 96% and 97% respectively.

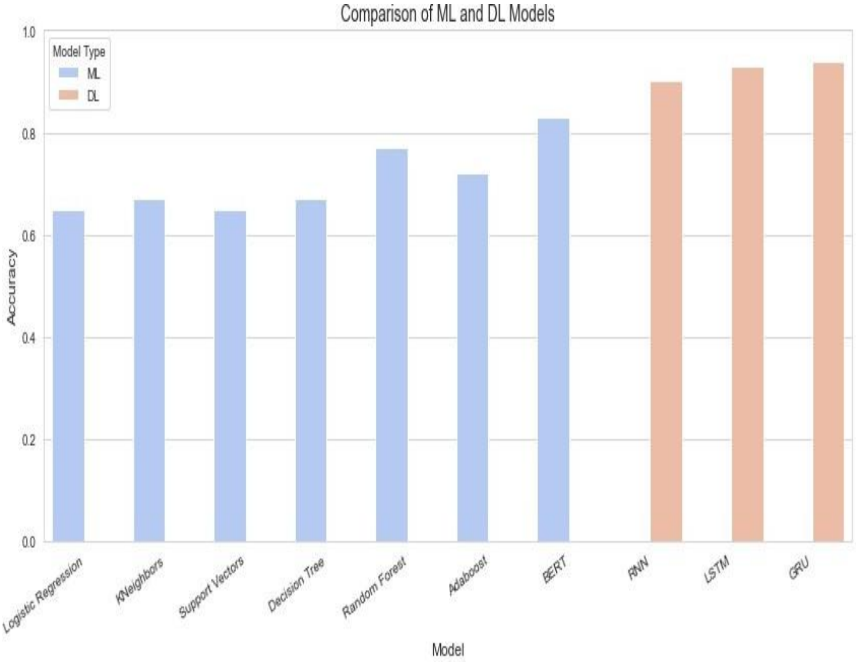

Fig. 7. Comparison graph of ML Algorithms and DL models.

This shows that the networks are more capable of processing sequential as well as large scale large dimensionality of data and temporal relation so more suitable for tasks with sequential or large data.

References

1. H. Utama, “Sentiment analysis in airline tweets us- ing mutual information for feature selection,” in 2019 4th International Conference on Information Tech- nology, Information Systems and Electrical Engineer- ing (ICITISEE), 2019, pp. 295–300. DOI: 10 . 1109 / ICITISEE48480.2019.9003903.
2. A. S. V. Akshada Sunil Shitole, “Machine learning based airlines tweets sentiment classification,” Inter- national Journal of Computer Applications, vol. 185, no. 20, pp. 32–35, Jul. 2023, ISSN: 0975-8887. DOI: 10 . 5120 / ijca2023922922. [Online]. Available: https ://ijcaonline.org/archives/volume185/number20/32810- 2023922922/.
3. L.-C. Cheng and S.-L. Tsai, “Deep learning for au- tomated sentiment analysis of social media,” in 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 1001–1004. DOI: 10.1145/3341161.3344821.
4. A. Rane and A. Kumar, “Sentiment classification system of twitter data for us airline service analysis,” in 2018 IEEE 42nd Annual Computer Software and Applications Conference

(COMPSAC), vol. 01, 2018, pp. 769–773. DOI: 10.1109/COMPSAC.2018.00114.

5. M. J. Adarsh and P. Ravikumar, “An effective method of predicting the polarity of airline tweets using sentimen- tal analysis,” in 2018 4th International Conference on Electrical Energy Systems (ICEES), 2018, pp. 676–679. DOI: 10.1109/ICEES.2018.8443195.
6. N. Raihen, S. Akter, F. Tabassum, F. Jahan, and S. Begum, “A statistical analysis of excess mortality mean at covid-19,” Computational Journal of Mathematical and Statistical Sciences, vol. 2, pp. 223–239, Sep. 2023. DOI: 10.21608/CJMSS.2023.229207.1014.
7. J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” CoRR, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http:// arxiv.org/abs/1810.04805.
8. Sharma, Harsh, Pangaonkar, Satyajit, Gunjan, Reena, and Rokade, Prakash, “Sentimental analysis of movie reviews using machine learning,” ITM Web Conf., vol. 53, p. 02 006, 2023. DOI: 10 . 1051 / itmconf / 20235302006. [Online]. Available: https://doi.org/10. 1051/itmconf/20235302006.
9. N. Kai Ning Loh, C. Poo Lee, T. Song Ong, and K. Ming Lim, “Mpnet-grus: Sentiment analysis with masked and permuted pre-training for language un- derstanding and gated recurrent units,” IEEE Access, vol. 12, pp. 74 069–74 080, 2024. DOI: 10 . 1109 / ACCESS.2024.3394930.
10. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning tech- niques,” EMNLP, vol. 10, Jun. 2002. DOI: 10 . 3115 /1118693.1118704.