**Introduction**

This study classifies language, feelings, perceptions, and emotions into good, negative, and neutral sentiments using sentiment analysis on tweets from airline passengers. Natural language processing (NLP) technique known as sentiment analysis is used to examine tweets regarding travel experiences, interactions with customer service, happiness, and general well-being. The report emphasizes the value of sentiment analysis to the airline industry, as it enables carriers to enhance the traveler experience and implement preventative measures. However, in order to make sentiment analysis more accurate and beneficial for the airline business, researchers must develop sophisticated natural language processing tools to comprehend online passenger impressions and feedback regarding airlines.

**Background of the Study**

Airlines are now able to measure customer satisfaction, pinpoint areas for improvement, and enhance the quality of their services by using sentiment analysis of their tweets. The airline sector prioritizes the needs of its customers, and millions of passengers share their thoughts and experiences on websites like X. Sentiment analysis is a subfield of Natural Language Processing (NLP) that categorizes textual information into three categories: positive, negative, and neutral perceptions. In the past, researchers have used a range of approaches and techniques to extract sentiment from textual data. These include deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), as well as traditional machine learning algorithms like Naive Bayes, Support Vector Machines (SVM), and Random Forests. A methodology utilizing word embedding with the Glove dictionary approach and the n-gram approach, incorporating Support Vector Machine, was developed by Kumar and Zymbler (2019), Convolutional Neural Network (CNN), Support Vector Machine (SVM), and other Artificial Neural Network (ANN) designs to classify tweets into positive and negative sentiment categories. After 2700 iterations on the validation set, CNN outperformed SVM and ANN models, obtaining an accuracy of 92.3%.

A study on sentiment analysis of airline tweets using machine learning approaches was carried out by Ravi Kumar et al. in 2021. They employed three methods: neural networks, Support Vector Machines (SVM), and decision trees. The neural network-based method outperformed SVM and decision trees, with the best accuracy of 75.99%. In order to improve sentiment accuracy, Aljedaani et al. (2022) presented a hybrid sentiment analysis strategy that combines deep learning models with lexicon-based techniques. They compared the original annotations to the classification accuracy of TextBlob, a lexicon-based technique. Additionally, they looked into the effectiveness and efficiency of the Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF) approaches. The findings indicated that TextBlob-assigned sentiments performed better than original sentiments, with the LSTM-GRU model exhibiting the greatest accuracy among the models.

In order to analyze the sentiment of tweets pertaining to airlines, Akhmad et al. (2023) compared a number of machine learning techniques. They separated the data into two stages and represented the features using the TF-IDF model. Class imbalance was addressed in the first phase using the Synthetic Minority Over-sampling Technique (SMOTE), and in the second phase using Stratified K-Fold Cross-Validation in the absence of SMOTE. With SMOTE oversampling applied, the Random Forest model produced the maximum accuracy, 97.56%, but without SMOTE, it was just 92.21%.

Samir et al. (2023) overcame difficulties including time-consuming surveys and inconsistent findings by using sentiment analysis to evaluate airline consumer feedback. They examined Skytrax Airline Customers' Feedback data using deep learning models such as Recurrent Neural Networks, Long Short-Term Memory, GRU, CONV1D, and BERT. Techniques for glove embedding were applied to improve sentiment classification. In a comparison research, LSTM outperformed other models in sentiment analysis tasks, achieving the maximum classification accuracy of 91%.

All in all, these research demonstrate the potential of sentiment analysis methods, such deep learning and machine learning, in examining airline tweets to extract insightful information about passenger feelings and improve the overall traveler experience. Airlines may successfully assess and respond to consumer feedback in real time by utilizing modern text analysis tools along with natural language processing (NLP). This approach ultimately leads to increased customer satisfaction and loyalty.

As suggested by the background of the study, this study aims to implement a hybrid approach incorporating lexicon-based methods and machine learning techniques. Specifically, the objective is to:

* Utilize lexicon-based approaches to categorize sentiment outcomes into negative, positive, and neutral categories.
* Employ machine learning algorithms for sentiment classification.
* Utilize deep learning techniques to further improve sentiment classification accuracy.
* Test the predictive and classification performance of the machine learning and deep learning models using performance metrics such as precision, recall, F1 scores, and accuracy.

This goal is specific because it provides a precise definition of tasks pertaining to airline tweet sentiment analysis. It is Measurable since certain measures will be used to assess the models' performance. Because it expands on already-established sentiment analysis approaches and procedures, the goal is attainable. It is Relevant because it discusses how the airline industry has to improve sentiment analysis in order to increase consumer satisfaction. Lastly, the study's objective is time-bound since it intends to finish the implementation and evaluation within a predetermined window of time.

**Data Preprocessing**

In order to do sentiment analysis, secondary data was collected during the data collecting phase from the Kaggle website (kaggle.com, n.d.). The project's chosen dataset consisted of tweets from passengers aboard airlines. The dataset has 14640 observations and fifteen (15) variables at first. However, other factors were eliminated during the data processing phase, leaving only important variables—like the unstructured content found in customer tweets—retained for additional examination.

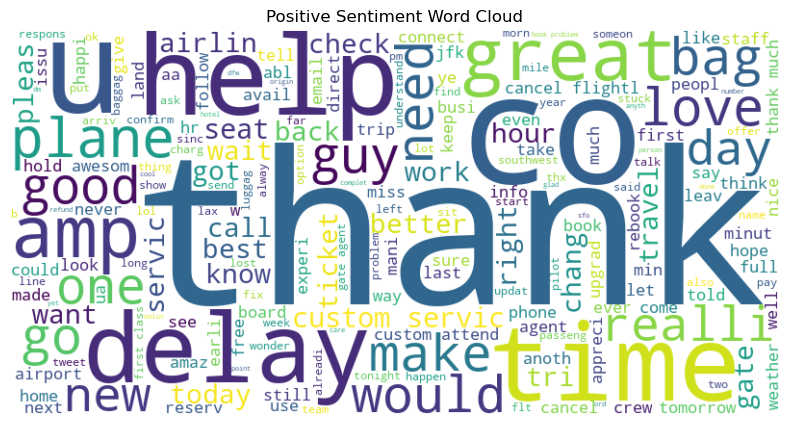
The data was cleaned by eliminating missing observations during the preprocessing and exploratory data analysis (EDA) stage. Text preparation techniques were then used to clean and standardize the textual content. This entails changing all punctuation to lowercase, tokenizing the text into a list of words, and getting rid of stopwords. Text normalization methods, such lemmatization and stemming, were used to reduce words to their root or base forms. Lexicon-based sentiment analysis was used for the exploratory data analysis following the preprocessing phases. Textblob was used for this challenge in order to determine the input text's sentiment polarity. Positive polarity scores were classed as neutral, negative polarity scores were classified as negative, and polarity scores equal to or higher than A score of zero was deemed positive. Figure 1 displayed the sentiment outcome's distribution, while Figures 2-4 displayed the sentiment outcome's word cloud.

The distribution of the sentiment classification presented in Figure 1 suggests that the distribution is balanced with 38% of the tweets classified as neutral, 37% as positive, and 25% as negative. According to the word cloud presented in Figure 4, the neutral sentiment encompasses words like routine inquiries and general information that have prevalent tones that are either positive or negative. The word cloud in Figure 1 displays keywords like "appreciation," "thanks," and "good customer service" which expresses favorable content. While Figure 3 displays terms like "delayed flight," and "poor customer service," which indicate customer displeasure.

A pie chart with different colors

Description automatically generated

*Figure 1: Pie chart displaying the distribution of the sentiment outcome.*



*Figure 2: Word cloud of the positive outcome.*

A close up of words

Description automatically generated

*Figure 3: Word cloud of the negative outcome.*

A close up of words

Description automatically generated

*Figure 4: Word cloud of the negative outcome.*

**AI Models**

**Traditional Machine Learning Models**

The five models selected for this section were chosen based on the reviewed literature while SVM and Random Forest classifier were chosen to achieve the aim of the study. The justification for choosing SVM and Random Forest was they predict sentiment outcomes with an accuracy greater than 80%. SVM can capture complex decision boundaries and handle high dimensional data which makes it effective in classifying textual data due to the dynamic nature of people's perceptions. Random Forest is robust to overfitting and can handle noise data such as negative perceptions.

1. Naïve Bayes

Naive Bayes, a probabilistic classifier rooted in Bayes' theorem with strong independence assumptions between features, finds widespread application in text classification tasks owing to its simplicity and efficiency (Akhmad et al., 2023). Its straightforward and computationally efficient nature makes it a preferred choice for handling text data (Akhmad et al., 2023). Although Naive Bayes may lack the capability to capture intricate relationships between words, it frequently delivers satisfactory performance in sentiment analysis tasks, particularly when confronted with limited training data (Akhmad et al., 2023).

1. Support Vector Machine

Support Vector Machines (SVM) is an efficient tool in supervised learning, specifically tailored for classification tasks (Awad et al., 2015). Justifying its application in sentiment analysis, SVMs showcase a remarkable capacity to handle high-dimensional data and delineate decision boundaries (Awad et al., 2015). Their proficiency extends to effectively classifying textual data by translating it into a higher-dimensional space, rendering them well-suited for NLP tasks characterized by non-linear relationships.

1. Logistic Regression

Logistic Regression, a linear model employed in binary classification tasks, operates by estimating probabilities through the logistic function and subsequently making predictions based on a predefined threshold (Wilson et al., 2015). This model is characterized by its simplicity yet effectiveness, particularly in sentiment analysis endeavors, where binary classification is common (Wilson et al., 2015). Its interpretability and compatibility with sparse text data render it suitable for various NLP applications, offering straightforward insights and reliable performance in sentiment analysis tasks.

1. Random Forest

Random Forest, an ensemble learning technique, is distinguished by its construction of multiple decision trees during training, whose predictions are amalgamated through a voting mechanism (Akhmad et al., 2023). This model's versatility and robustness make it a standout performer in sentiment analysis tasks, especially when confronted with high-dimensional text data (Akhmad et al., 2023).

1. Decision Trees

According to Patel et al. (2023), a decision tree is a supervised learning method that divides data into subsets based on feature values. The resulting tree-like structure has internal nodes that represent decisions based on features and leaf nodes that represent class labels. Decision trees are a useful tool in sentiment analysis that help find and classify patterns in textual data. They are appropriate for NLP jobs where comprehending classification decisions is critical because to their interpretability and capacity to handle numerical and categorical input (Patel et al., 2023).

**Deep Learning Models**

This study adopts LSTM and BiLSTM because of the high accuracy tendency they possess in classifying sentiment outcomes. BiLSTMs offer bidirectional processing, capturing dependencies in both directions for a comprehensive understanding of sentence sentiment (You et al., 2022; Samir et al. 2023). This enhances the model's ability to capture complex relationships within the text, crucial for tasks requiring complete context. In contrast, LSTMs excel at retaining information over extended sequences, ideal for understanding sentiment within sentences (Samir et al., 2023). While lacking bidirectional capability, LSTMs remain efficient and robust for sentiment analysis tasks, especially in scenarios where bidirectional context is unnecessary or computational resources are limited (You et al., 2022). Other deep-learning suggestions are:

1. Recurrent Neural Networks (RNN):

RNNs are neural networks with internal states and memory that are intended to operate with sequential data. With time, they can identify patterns and dependencies (Das et al., 2023). Because RNNs can parse word or character sequences, they are frequently utilized in NLP applications. They do exceptionally well in tasks requiring sequential relationships and context, such as sentiment analysis, machine translation, and language modeling (Das et al., 2023).

1. Gated Recurrent Units (GRU):

GRUs are another variant of RNNs designed to address the vanishing gradient problem. They have fewer parameters compared to LSTMs but can still capture long-term dependencies in sequences (Samir et al., 2023). GRUs offer a balance between simplicity and effectiveness for NLP tasks. They can handle sequential data efficiently and are suitable for sentiment analysis tasks where memory and context preservation are important (Samir et al., 2023).

1. Convolutional Neural Networks (CNN):

CNNs are well-liked deep learning models for image identification-related applications. They consist of pooling layers, which are employed to extract hierarchical features from input data, following convolutional layers (Aljedaani et al., 2022). Although their primary use is in image processing, CNNs can also be applied to natural language processing (NLP) tasks including text categorization and sentiment analysis (Aljedaani et al., 2022).

**Implementation**

During the implementation stage, NumPy, Pandas, Matplotlib, Seaborn, NLTK, TextBlob, Scikit-learn, and Keras were among the Python libraries used. The preprocessed text data was used to create TF-IDF (Term Frequency-Inverse Document Frequency) features after the dataset was divided into 80% training and 20% testing sets.The implementation process involves the utilization of traditional machine learning and deep learning approaches for sentiment analysis of airline tweets. The ML models were trained using RandomForestClassifier and SVC from the Scikit-learn library and their hyperparameters were fine-tuned using GridSearchCV to optimize their predictive and classification performance.

On the other hand, for deep learning models, the text data is tokenized and padded to ensure uniform input length. Two types of models, LSTM and BiLSTM, were explored, with various combinations of hyperparameters such as embedding dimension, LSTM units, and dropout rates. These models were trained on the preprocessed training data, and their performance was evaluated based on validation accuracy. The model configuration with the highest validation accuracy was selected as the best model, and its performance was assessed on the testing set using metrics like confusion matrix and classification report.

The choice of traditional machine learning models like RandomForestClassifier and SVC is justified by their simplicity, interpretability, and effectiveness in handling tabular data. Hyperparameter tuning using GridSearchCV allows for optimizing the models' performance. Meanwhile, deep learning models such as LSTM and BiLSTM are suitable for processing sequential data like text and capturing complex patterns and dependencies within the data. The hyperparameter tuning process ensures that these models are fine-tuned to achieve the best performance, making the implementation comprehensive and robust for sentiment analysis of airline tweets.

**Evaluation and Testing**

For this task, the performance metrics were utilized to test and compare the predictive and classification performance of the selected models. In comparing the performance of the models, the model that has precision, recall, F1 score, and accuracy close to 1 (100%) was selected as the best model in classifying airline customer tweets.

*Table 1: Evaluation Metrics Utilized*

|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Formula (Confusion Matrix)** |
| Precision | Proportion of true positive predictions out of all positive predictions made by the model | Precision = TP / (TP + FP) |
| Recall | The proportion of true positive instances correctly identified by the model | Recall = TP / (TP + FN) |
| F1-score | The harmonic mean of precision and recall provides a balance between the two metrics | F1-score = 2 \* (Precision \* Recall) / (Precision + Recall) |
| Accuracy | Overall correctness of the model's predictions | Accuracy = (TP + TN) / (TP + TN + FP + FN) |

The predictive comparison of the four models is presented in Table 2. In predicting the negative sentiment instances correctly, the F1 score suggests that the RandomForest model achieved 0.7 (70%), SVM achieved 0.74, LSTM didn't classify any instances correctly, and BiLSTM achieved 0.73. In terms of predicting the neutral cases correctly, SVM outperformed the other models with a score of 0.84, followed by BiLSTM at 0.82, RandomForest at 0.83, and LSTM at 0.56. In predicting positive sentiment cases, RandomForest and BiLSTM both achieved an F1-score of 0.8, SVM achieved 0.79, and LSTM didn't classify any instances correctly. In terms of overall accuracy, SVM performed the best at 0.8039, followed by RandomForest at 0.7927 and BiLSTM at 0.79. LSTM, however, showed significantly lower accuracy at 0.39. Overall, SVM demonstrated the highest overall accuracy among the models, while RandomForest and BiLSTM showed competitive performance across all sentiment classes based on the F1-score. However, the LSTM model performed poorly compared to the others, especially in terms of F1-score and accuracy. This result is in line with Rustam et al. (2019). Also, in a study by Karaman et al. (2022), SVM with an accuracy of 98.0% outperformed the LSTM model (97%) and CNN-RNN (96%).

*Table 2: Comparison of the Predictive Performance of the Selected Models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | RandomForest | SVM | LSTM | BiLSTM |
| Precision (Class 0) (Negative) | **0.82** | 0.79 | 0 | 0.72 |
| Recall (Class 0) (Negative) | 0.62 | 0.69 | 0 | **0.74** |
| F1-score (Class 0) (Negative) | 0.7 | **0.74** | 0 | 0.73 |
| Precision (Class 1) (Neutral) | 0.77 | 0.8 | 0.39 | **0.83** |
| Recall (Class 1) (Neutral) | 0.89 | 0.89 | **1** | 0.82 |
| F1-score (Class 1) (Neutral) | 0.83 | **0.84** | 0.56 | 0.82 |
| Precision (Class 2) (Positive) | 0.81 | **0.82** | 0 | 0.8 |
| Recall (Class 2) (Positive) | **0.8** | 0.79 | 0 | 0.79 |
| F1-score (Class 2) (Positive) | **0.8** | 0.79 | 0 | **0.8** |
| Accuracy | 0.7927 | **0.8039** | 0.39 | 0.79 |

**Conclusion**

Based on the findings, it is recommended that airline sectors consider implementing sentiment analysis models, particularly those based on Support Vector Machines (SVM), Random Forest, and BiLSTM, for analyzing customer tweets. These models demonstrate strong performance in accurately classifying sentiment, which can provide valuable insights into customer satisfaction levels, concerns, and overall sentiment toward airline services. By leveraging sentiment analysis, airlines can promptly identify and address customer issues like delays and poor service and enhance overall customer experience.

**References (Havard Referencing Style)**

1. Akhmad, E.P.A., Adi, K. and Widodo, A.P., 2023. Machine learning approach to customer sentiment analysis in twitter airline reviews. In E3S Web of Conferences (Vol. 448, p. 02044). EDP Sciences.
2. Aljedaani, W., Rustam, F., Mkaouer, M.W., Ghallab, A., Rupapara, V., Washington, P.B., Lee, E. and Ashraf, I., 2022. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Knowledge-Based Systems, 255, p.109780.
3. Awad, M., Khanna, R., Awad, M. and Khanna, R., 2015. Support vector machines for classification. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pp.39-66.
4. Barakat, H., Yeniterzi, R. and Martin-Domingo, L., 2021. Applying deep learning models to twitter data to detect airport service quality. *Journal of air transport management*, *91*, p.102003.
5. Bigne, E., Andreu, L., Hernandez, B. and Ruiz, C., 2018. The impact of social media and offline influences on consumer behaviour. An analysis of the low-cost airline industry. *Current Issues in Tourism*, *21*(9), pp.1014-1032.
6. Das, S., Tariq, A., Santos, T., Kantareddy, S.S. and Banerjee, I., 2023. Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. Machine Learning for Brain Disorders, pp.117-138.
7. Devika, M.D., Sunitha, C. and Ganesh, A., 2016. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, *87*, pp.44-49.
8. Hasib, K.M., Habib, M.A., Towhid, N.A. and Showrov, M.I.H., 2021, February. A novel deep learning based sentiment analysis of twitter data for us airline service. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (pp. 450-455). IEEE.
9. kaggle.com. (n.d.). Sarvesh Kaushik US Airlines Sentiment (ULMFIT). [online] Available at: <https://www.kaggle.com/code/pantanmol/sarvesh-kaushik-us-airlines-sentiment-ulmfit/input>.
10. Karaman, Y., Akdeniz, F., Savaş, B.K. and Becerikli, Y., 2022, October. A comparative analysis of svm, lstm and cnn-rnn models for the bbc news classification. In *The Proceedings of the International Conference on Smart City Applications* (pp. 473-483). Cham: Springer International Publishing.
11. Kumar, S. and Zymbler, M., 2019. A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1), pp.1-16.
12. Patel, A., Oza, P. and Agrawal, S., 2023. Sentiment analysis of customer feedback and reviews for airline services using language representation model. Procedia Computer Science, 218, pp.2459-2467.
13. Ravi Kumar, G., Venkata Sheshanna, K. and Anjan Babu, G., 2021. Sentiment analysis for airline tweets utilizing machine learning techniques. In International Conference on Mobile Computing and Sustainable Informatics: ICMCSI 2020 (pp. 791-799). Springer International Publishing.
14. Rustam, F., Ashraf, I., Mehmood, A., Ullah, S. and Choi, G.S., 2019. Tweets classification on the base of sentiments for US airline companies. *Entropy*, *21*(11), p.1078.
15. Samir, H.A., Abd-Elmegid, L. and Marie, M., 2023. Sentiment analysis model for Airline customers’ feedback using deep learning techniques. International Journal of Engineering Business Management, 15, p.18479790231206019.
16. Wilson, J.R., Lorenz, K.A., Wilson, J.R. and Lorenz, K.A., 2015. Standard binary logistic regression model. Modeling binary correlated responses using SAS, SPSS, and R, pp.25-54.
17. You, H., Yang, J. and Xu, Z., 2022. Sentiment Analysis Based on CNN and BILSTM Model. In International Conference in Communications, Signal Processing, and Systems (pp. 175-182). Singapore: Springer Nature Singapore.