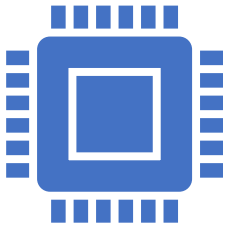
Airline Quality Reviews-Sentiment Analysis

Presented by: Simidola Lawani

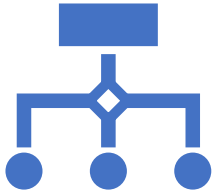
Date: 17/06/2024

Task 1: Data



Objective: To analyse airline review sentiments using advanced NLP techniques.

Key Steps: Data preprocessing, sentiment labeling, visualizations, modeling using GloVe embeddings, and LSTM.



Scrapped data from Skytrax website using beautifulsoup package in python after installing and importing useful packages.

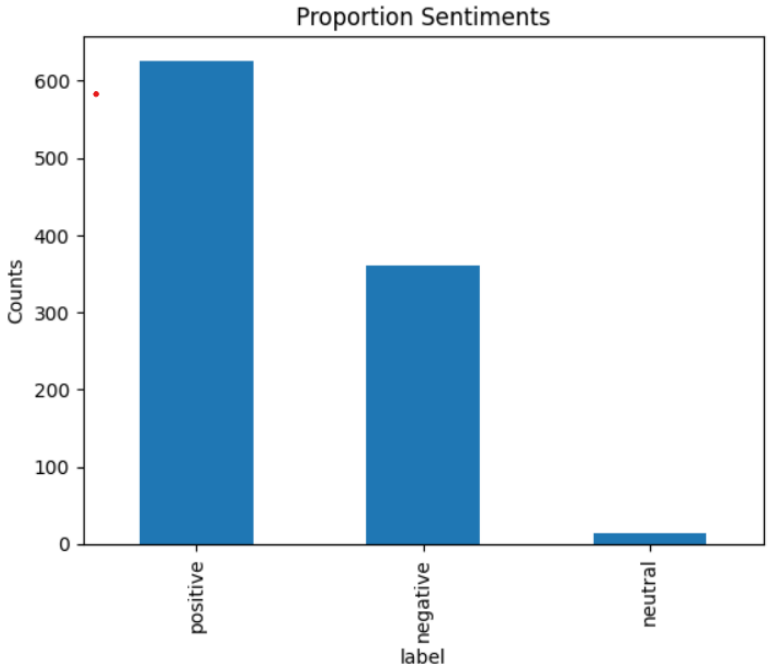
Number of data entries: 1000

Content of data: Text data of airline reviews.

Some important packages used for this project: Numpy, Pandas, Matplotlib, Seaborn, Beautifulsoup, Scikit-learn for Logistic regression, multinomial NB, Random Forest, Tensorflow LSTM, NLTK, Textblob etc

Task 2: Data Preprocessing

- Removed stopwords
- Removed HTML tags, URLs, and non-alphanumeric characters from the reviews. We do that with the help of the remove_tags function, and Regex functions are used for easy string manipulation by applying tokenization.
- Removed unwanted words like Trip Verified and other unwanted characters.
- Used Textblob to apply labelling.(Sentiments)

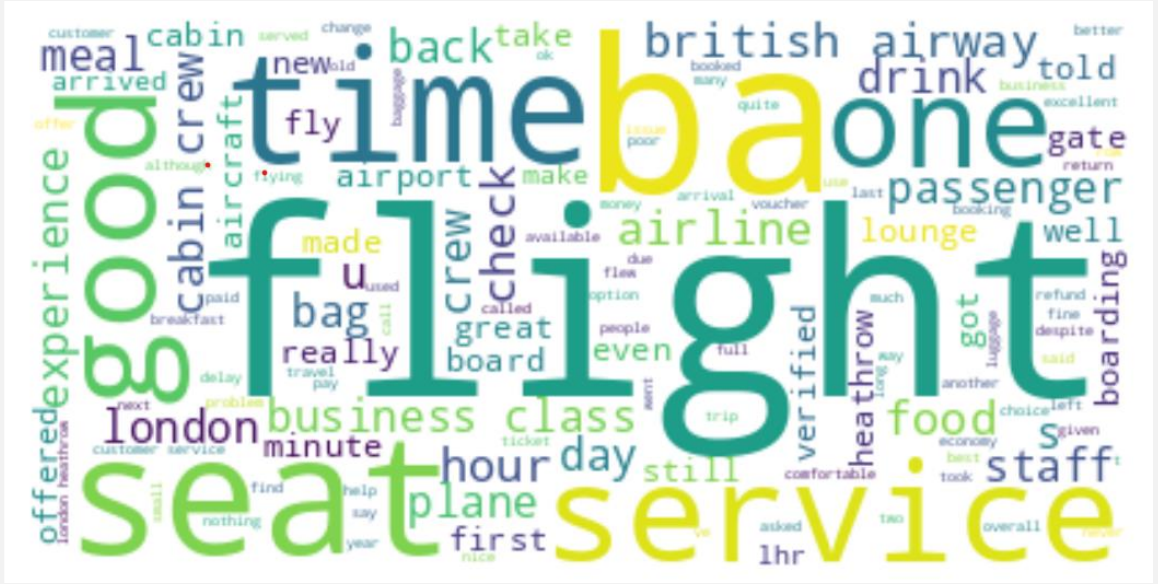


Task 3: Data Visualisation: Sentiments

- From the visualization we can observe we have more positive reviews than negatives. A very small portion of neutral reviews were discovered.

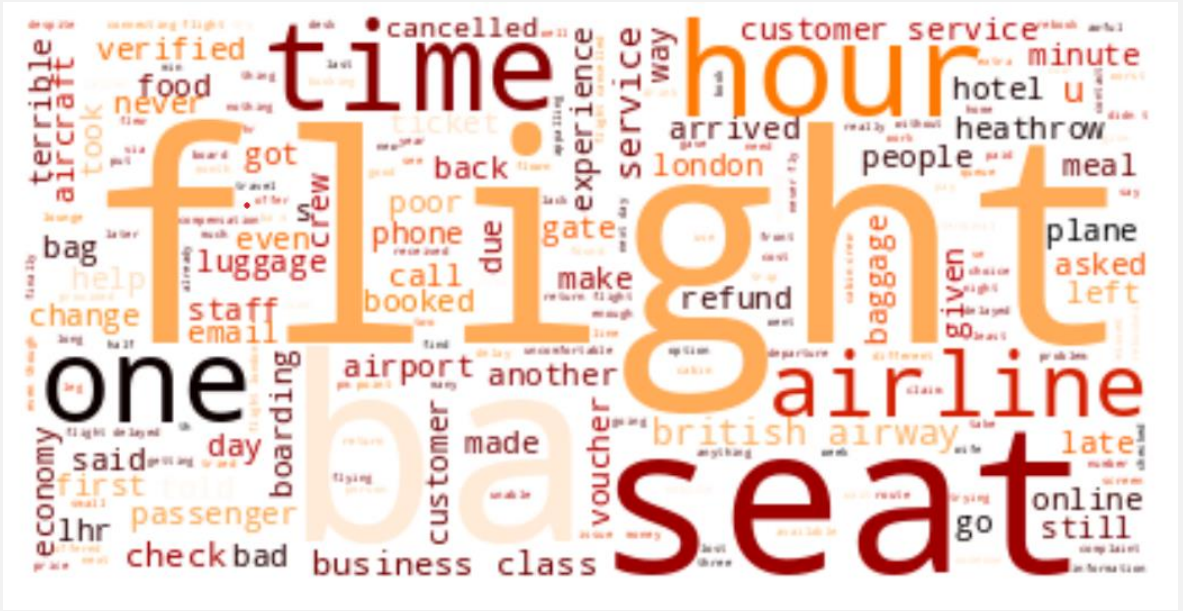
- The airline needs to do something about the negative reviews as it is not a good on the airline.

Data Visualisation: Positive Reviews from WordcLoud



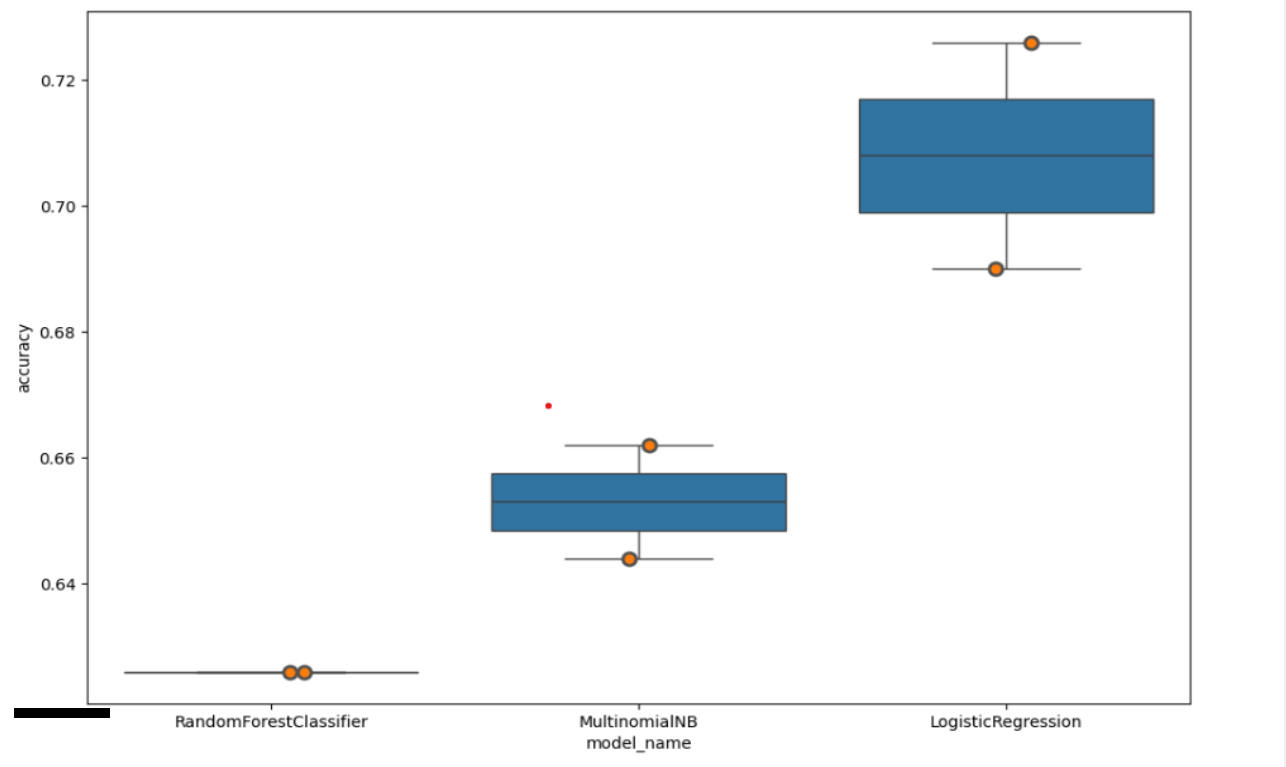
- Words like food, cabin crew, drink, staff, seat, good, time, service etc., appear in the positive reviews which is pointer to what is working for the airline.

Data Visualisation: Negative Reviews from WordcLoud

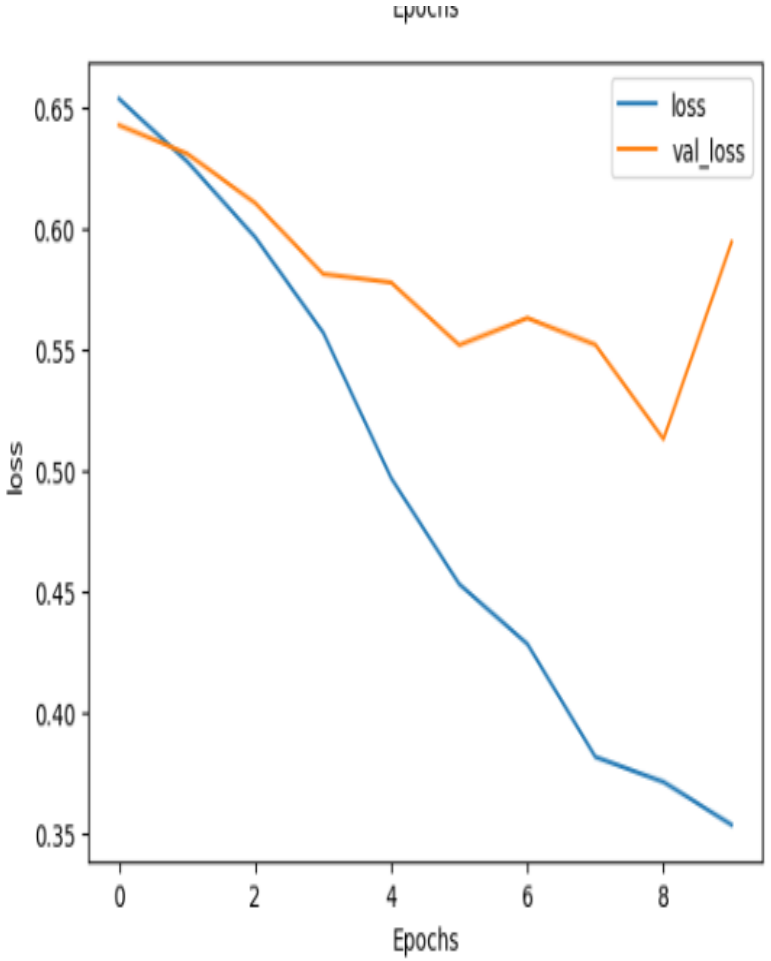


- Words like, cancelled, refund, seat, hour, boarding, terrible etc. These words reflect areas that the airline needs to improve on.

Modelling the Text Data



- Using three scikit learn algorithms: Random Forest, MultinomialNB and Logistic Regression. In terms of accuracy logistic regression out performed the other models with an accuracy of 70.8%.



Modelling: LSTM

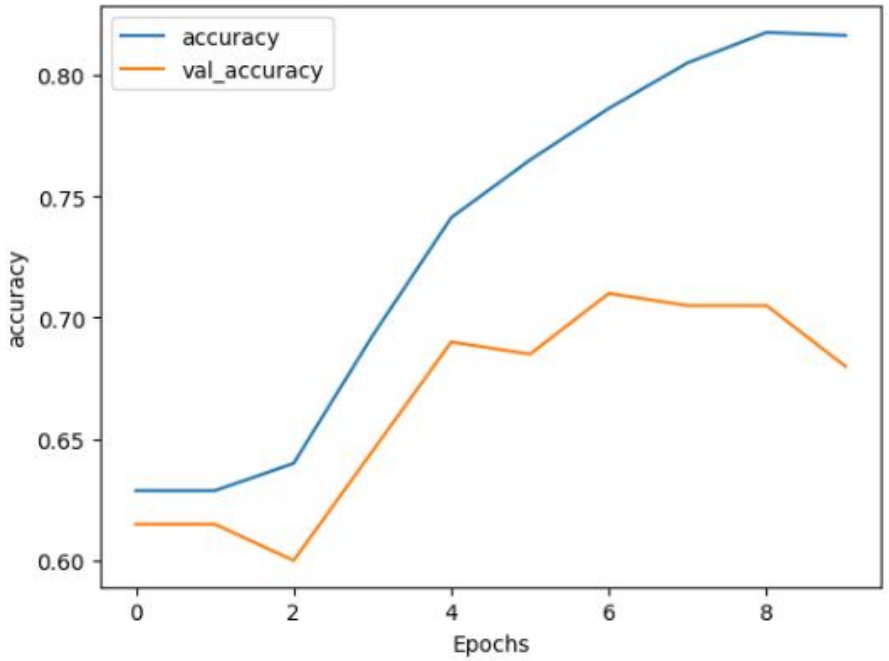
The grapgh shows the training loss (‘loss’) and validation loss (‘val_loss’) of a machine learning model over a number of epochs.

The **blue line** represents the training loss, which decreases as the number of epochs increases. This suggests that the model is learning and improving its performance on the training data over time.

The **orange line** represents the validation loss, which initially decreases but starts to increase after the fourth epoch. This could indicate that the model is starting to overfit to the training data, meaning it’s learning patterns specific to the training data that don’t generalize well to new data.

The increasing validation loss suggests that it might not be generalizing well to new data. Collecting more diverse training data may be valuable.

Modelling: LSTM



The graph shows the training accuracy (‘accuracy’) and validation accuracy (‘val_accuracy’) of the LSTM deep learning model over 10 epochs.

The **blue line** represents the training accuracy, which increases as the number of epochs increases. This suggests that the model is learning and improving its performance on the training data over time.

The **orange line** represents the validation accuracy, which also increases but has fluctuations and does not reach as high as the training accuracy by epoch 8. This suggests that while the model is learning from the training data, its performance on validation data is not improving at the same rate and even drops at certain points.

Adding more diverse training data will help with this issue to avoid overfitting as the data entries used is 1000 in number.

CONCLUSION

Summary: Successful sentiment analysis of airline reviews using NLP techniques.

Future Work:

Collect more diverse data to improve the model's generalization in predicting new dataset.

Explore other models like BERT or transformer-based models.

Apply to other domains for sentiment analysis.

P.S I was not able to attach my Jupyter Notebook for your perusal and I would have loved to