Enhancing Airline Sentiment Analysis through Classification: A Comparative Study of Logistic Regression, Decision Tree Classifiers, and Random Forest Models

Author: Anirban Das.

Date: 11th February, 2024.

Introduction

In the ever-evolving landscape of the airline industry, customer sentiments play a pivotal role in shaping perceptions and influencing business outcomes. This study delves into the intricate world of airline sentiment analysis, focusing on the application of unsupervised learning using VADER scoring techniques. By leveraging this technique, the objective is to decipher the nuanced emotions expressed in customer feedback, allowing airlines to glean actionable insights for refining their services, addressing pain points, and ultimately elevating the overall passenger experience.

In addition to exploring sentiments within airline customer reviews, this study employs supervised classification methods such as Logistic Regression, Decision Tree Classifiers, and Random Forest models to predict and classify sentiments. The reviews undergo transformation using TF-IDF (Term Frequency-Inverse Document Frequency) scoring, a technique that captures the significance of words in a document relative to the entire corpus. By integrating these algorithms with TF-IDF scoring, the research seeks not only to analyse existing sentiments but also to predict and classify sentiments in 'unseen' reviews. This dual approach aims to provide a comprehensive understanding of customer sentiments in the airline industry and to equip airlines with predictive tools for proactive service improvement. This exploration at the intersection of machine learning and customer sentiment aims to contribute to the ongoing pursuit of excellence in the aviation sector.

Data preprocessing and unsupervised sentiment analysis

Before starting with the supervised machine learning algorithms, the unsupervised techniques are used to assign the sentiment scores and tags which is further used for supervised machine learning.

Data extraction and cleaning

The data is downloaded from Kaggle https://www.kaggle.com/datasets/juhibhojani/airline-reviews/data

The data is consisting of twenty unique columns with above twenty-three thousand rows. After initial overview, the data is checked for null and duplicate values. Only the data related to verified reviews are selected, narrowing down the data to just above twelve thousand rows and with columns "*Airline Name*", "*Overall_Rating*", "*Review_Title*", "*Review*", "*Recommended*".

Data preparation and VADER scoring

The data is then prepared for the VADER (Valence Aware Dictionary and sEntiment Reasoner) scoring. This technique quantifies the sentiment of a piece of text by assigning a compound score that represents the overall sentiment intensity. The VADER scoring technique is based on a pre-built lexicon that includes both standard English words and domain-specific terms.

A function `def clean_data(dataframe)` is built to take any text as input and process the text by removing special characters, single characters with space, single characters from start, unnecessary spaces and punctuations. The text is further lower-cased, tokenized and the stop words are removed, finally joining the tokens back for VADER scoring.

The VADER scoring is implemented using the `SentimentIntensityAnalyzer` class from the Natural Language Toolkit `nlkt.sentiment`. The scoring technique is used to evaluate the review title and the reviews, then use the corresponding compound_score to classify the scores above zero as positive and below zero as negative, saving them to “Vader_Review” and “Vader_Review_Title” columns.

Most of the times it is very common to have a different title of the review than the actual review. So after assigning the VADER scores, classifying the review titles and the reviews, the mis-match is calculated. Out of 12319 verified opinions, there is around 30% mis-match of the title vs reviews.

Exploratory Data Analysis

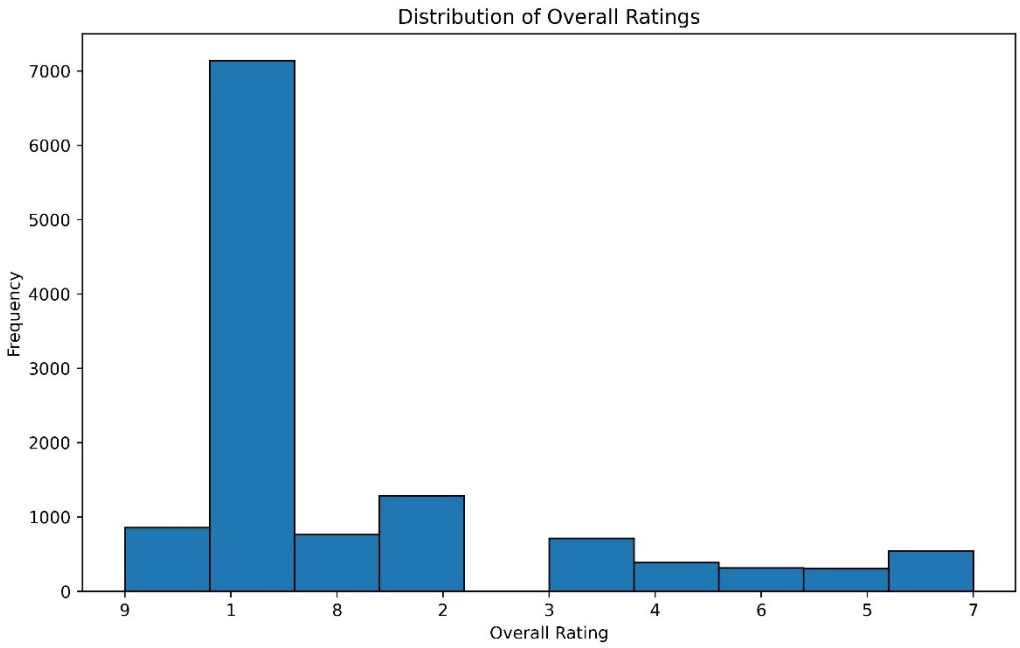

fig 1: Distribution of overall rating

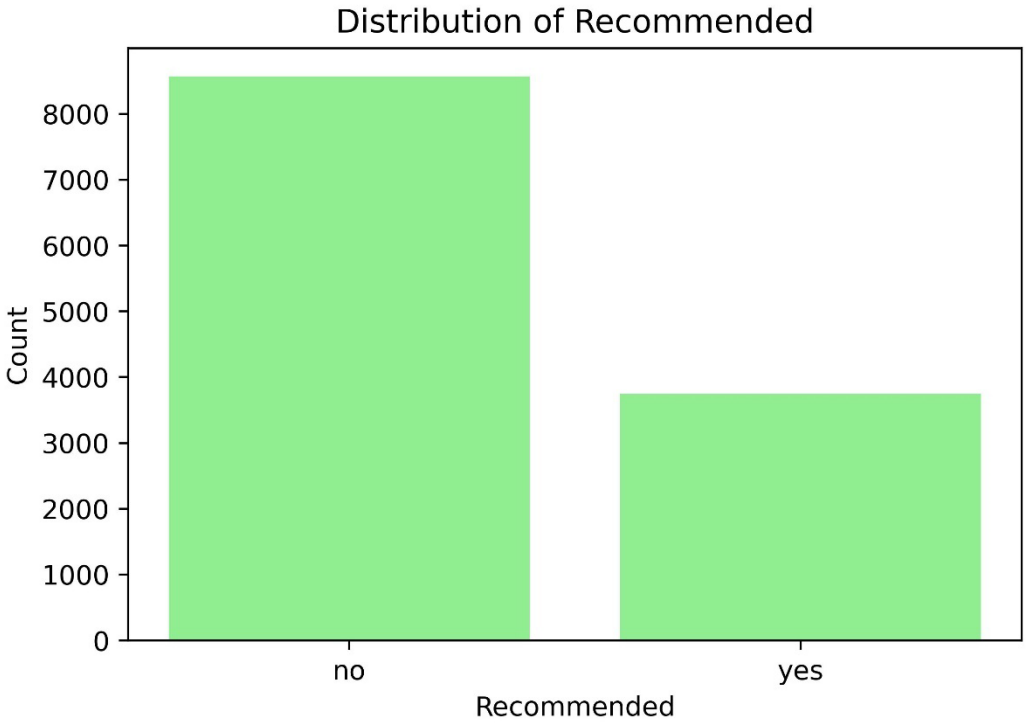

fig 2: Distribution of airlines being recommended

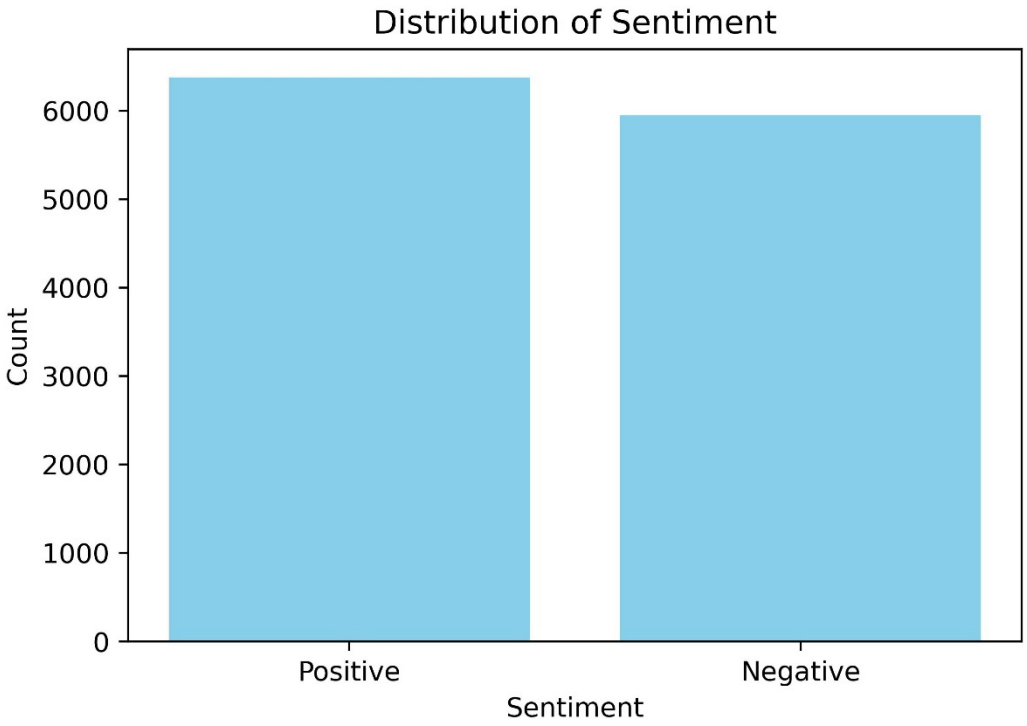

fig 3: Distribution of airlines sentiment

Classification Algorithms

After the unsupervised analysis is completed by scoring the reviews and assigning them into Positive and Negative sentiment, the next process is initiated for the comparative study of Sentiment Analysis using the Logistic Regression, Decision Tree Classifiers, and Random Forest Models.

Data preprocessing using TF-IDF

As the computer would not understand the textual data, the Reviews cannot be used as simple English language. Thus it is necessary to convert the reviews.

Here the Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer will be used for vectorizing the reviews into simple vectors of numbers based on the frequency of unique occurrences. This will be the explanatory variable for the classification task. The "clean" reviews after removing the stop words and preprocessing is used for the TF-IDF vectorizer as unclean data will affect the technique.

Simultaneously, the sentiments (Positive or Negative) are On-Hot Encoded to 1 and 0. Combining the TF-IDF vectors from Reviews and One-Hot Encoded sentiments, the features for classification task is prepared.

Logistic Regression Classifier

```
Accuracy: 0.78
Precision: 0.78
Recall: 0.78
F1 Score: 0.78
                  precision    recall  f1-score   support

           0          0.76      0.80      0.78      1779
           1          0.80      0.76      0.78      1917

accuracy                               0.78      3696
macro avg             0.78      0.78      0.78      3696
weighted avg          0.78      0.78      0.78      3696
```

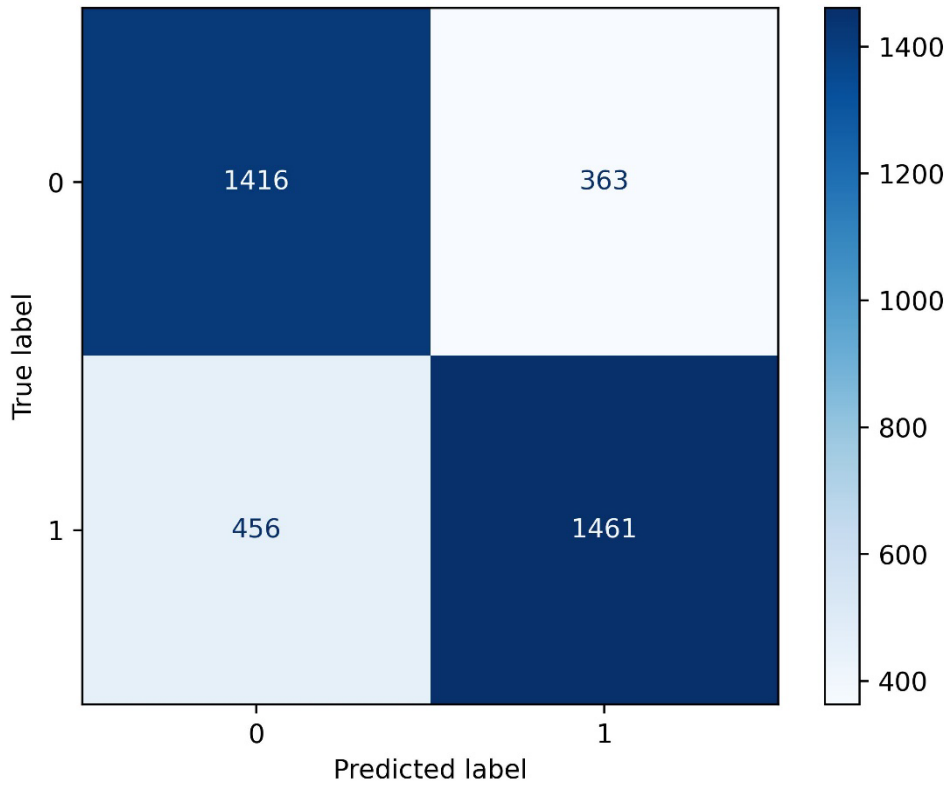

fig 4: Logistic Regression Confusion Matrix

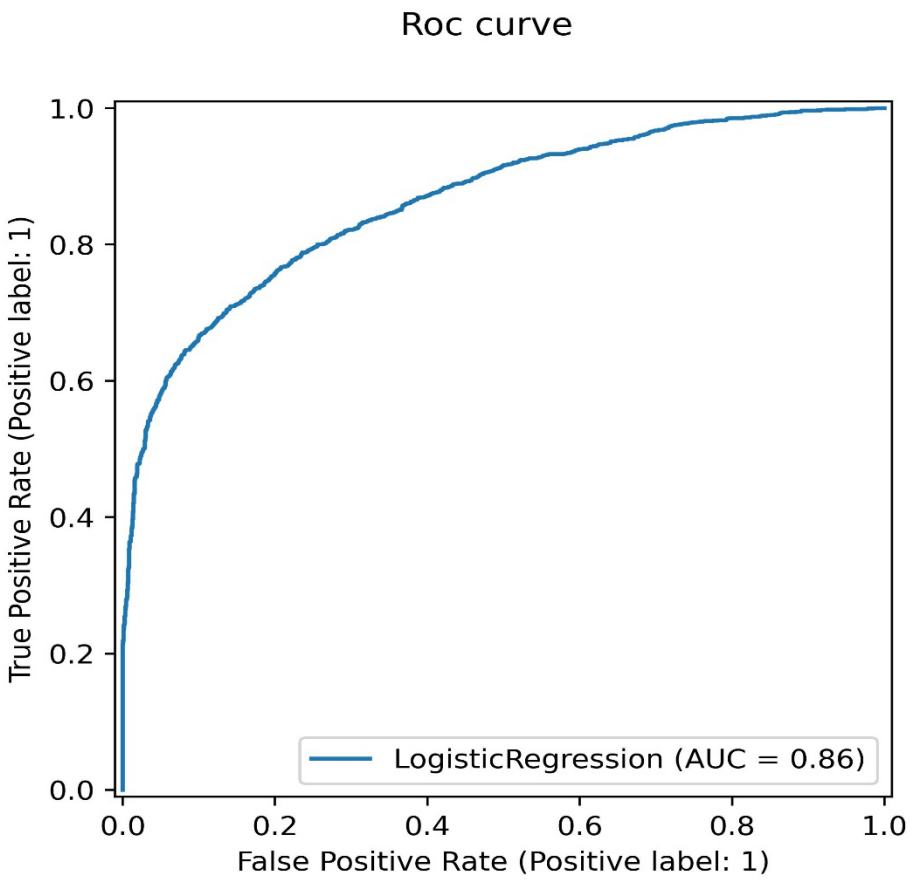

fig 5: Area Under the ROC Curve

From the above figures, the Logistic Regression showed a 78% accuracy with the area under the ROC curve being 0.86. The model correctly predicted the positive class (1) 1461 times (True Positives), the negative class (0) 1416 times (True Negatives), but incorrectly predicted the positive class 363 times (False Positives) and incorrectly predicted the negative class 456 times (False Negatives).

Decision Tree Classifier

```
Accuracy: 0.72
Precision: 0.72
Recall: 0.72
F1 Score: 0.72
                  precision    recall  f1-score   support

           0          0.72      0.71      0.71      1779
           1          0.73      0.74      0.73      1917

accuracy                               0.72      3696
macro avg             0.72      0.72      0.72      3696
weighted avg          0.72      0.72      0.72      3696
```

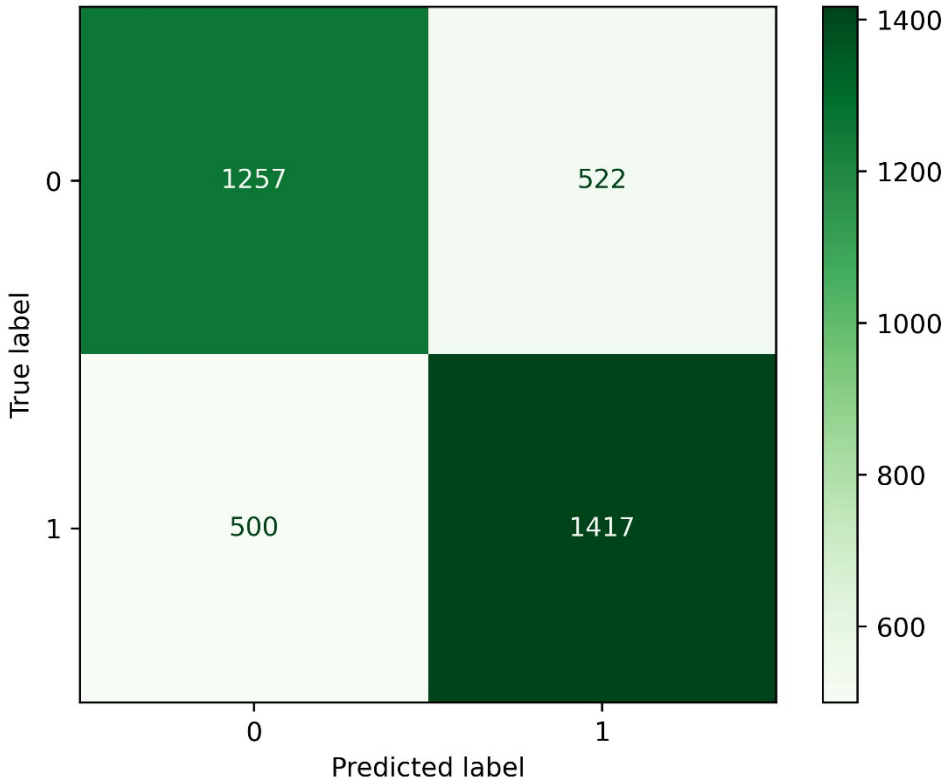

fig 6: Decision Tree Confusion matrix

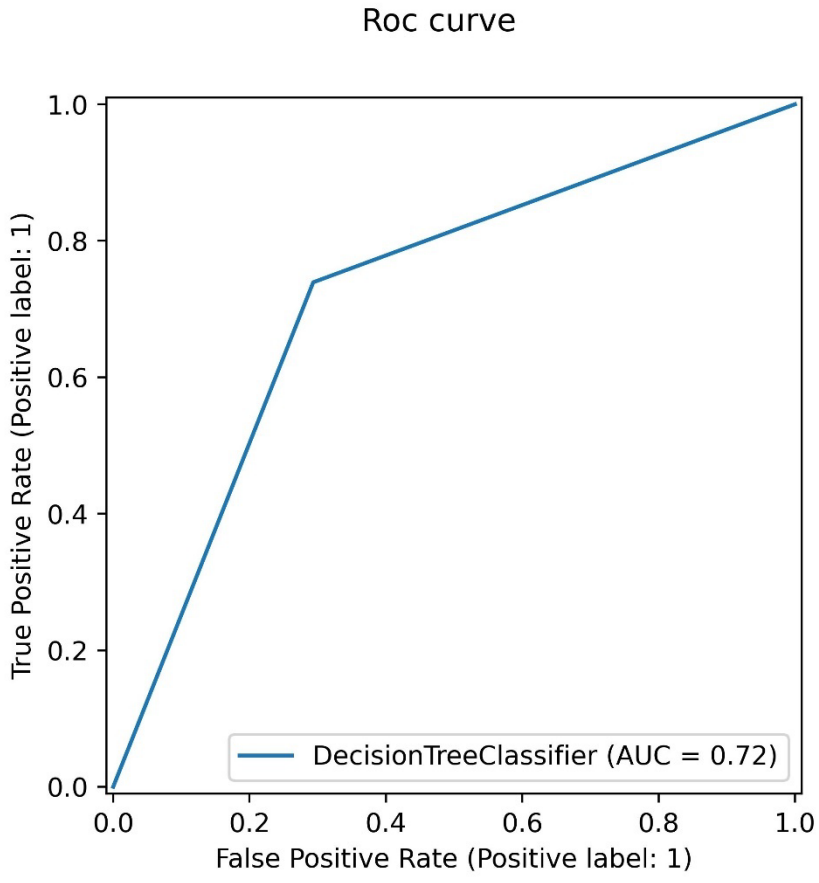

fig 7: Area under the ROC curve

From the above figures, the Decision Tree Classifiers showed a 72% accuracy with the area under the ROC curve being 0.72. The model correctly predicted the positive class (1) 1257 times (True Positives), the negative class (0) 1417 times (True Negatives), but incorrectly predicted the positive class 522 times (False Positives) and incorrectly predicted the negative class 500 times (False Negatives).

Random Forest

```
Accuracy: 0.80
Precision: 0.81
Recall: 0.80
F1 Score: 0.80
                  precision    recall  f1-score   support

           0          0.76      0.86      0.81      1779
           1          0.85      0.75      0.80      1917

accuracy                               0.80      3696
macro avg             0.81      0.81      0.80      3696
weighted avg          0.81      0.80      0.80      3696
```

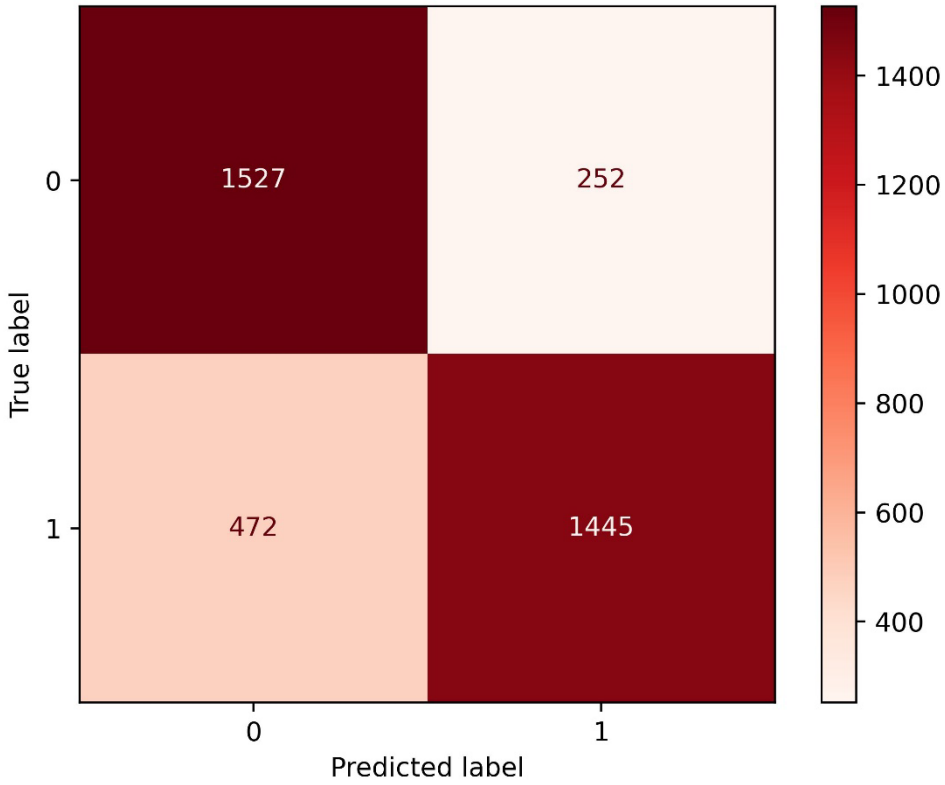

fig 8: Random Forest Confusion Matrix

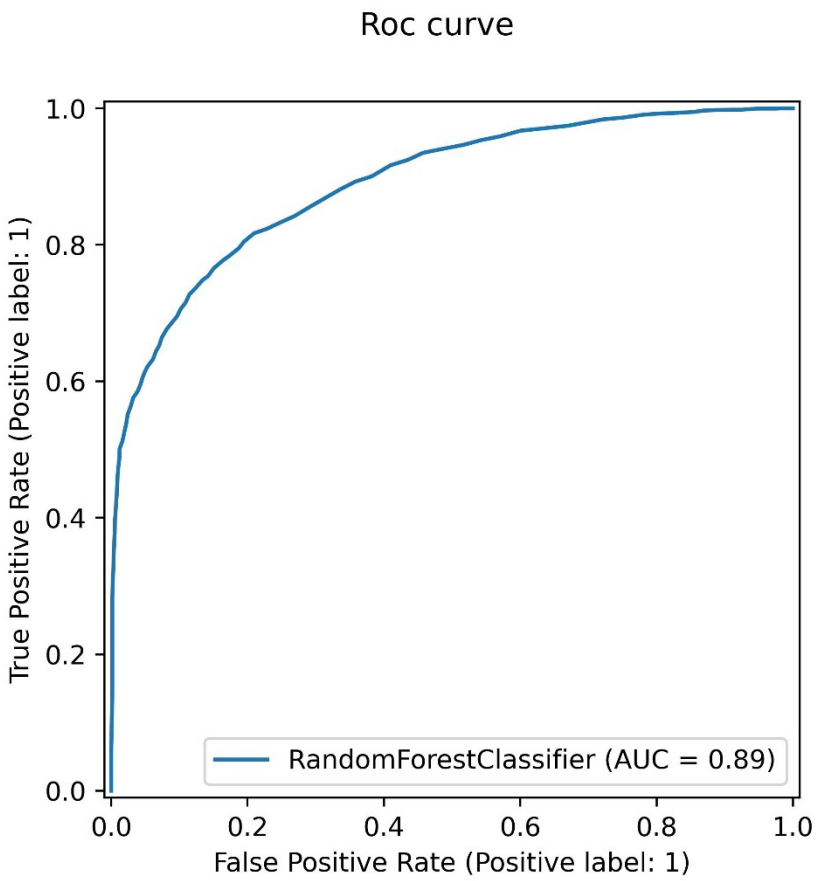

fig 9: Area under the ROC curve

From the above figures, the Random Classifiers showed a 80% accuracy with the area under the ROC curve being 0.89. The model correctly predicted the positive class (1) 1527 times (True Positives), the negative class (0) 1445 times (True Negatives), but incorrectly predicted the positive class 252 times (False Positives) and incorrectly predicted the negative class 472 times (False Negatives).

Conclusion

From the comparative study, the Random Forest classifier demonstrated superior performance with an 80% accuracy and an area under the ROC curve of 0.89. It exhibited notable true positives and true negatives, showcasing its robust predictive capabilities. On the other hand, the Decision Tree and Logistic Regression models achieved lower accuracies of 72% and 78%, respectively. This study can be further enhanced, by implementation of the deep learning algorithms, model tuning and exploring the ensemble methods. The application of real-time sentiment analysis and continuous model adaptation would ensure the models stay relevant in dynamic linguistic contexts.