Sentiments classification for US airline companies

15 December 2021

Abstract

Nowadays, social media produces a high quantity of textual content that grows exponentially. As they become the main communication channel, accessed by all segments of society. For instance, Twitter records more than 500 million tweets per day. Therefore, such platforms are the floor to express opinions, share feelings , build and develop relationships. Either by posts, comments, likes or direct messages. Therefore, with the availability of such text data through twitter API, we build multiple classification machine learning models which are Random Forest, Logistic Regression and Adaptive boosting. We focus on predicting customer satisfaction about airplane services in the US. And we adapt confusion matrix, recall and precision as evaluation metrics. The results show that Logistic regression is more accurate in terms of recall in predicting negative sentiments.

Design

Customer feedback is a very important player in the airline industry because it helps companies to improve their services. One of the traditional methods to achieve this goal is to ask for customer satisfaction by questionnaires or forms. These procedures are time consuming and require a lot of resources. Moreover, the results are inconsistent and inaccurate because not all customers will fill these forms seriously. On the other hand, Twitter produces a high quantity of textual content which mostly represents customers' opinions. Therefore, we can explore data science techniques in order to analyse this data and build classification models. Those new data-based approaches will give the airplanes companies the ability to predict customers' feedback directly from their tweets. In this project we use an available dataset from kaggle which contains tweets along with correspondent sentiments in order to build multiple classification models.

Data

The data set we used is a twitter data set available at : https://www.kaggle.com/crowdflower/twitter-airline-sentiment . It is released by CrowdFlower and has a total of 14640 tweets. Tweets of six major US Airlines have evolved:US Airways,Southwest, United, Virgin America, Delta. The tweets are labeled as negative, positive and neutral sentiment. Thus, the ability to apply classification models.

Algorithms

In this section we will describe the main approaches we use to prepare the data used for sentiment analysis and the proposed models to build sentiment classification models.

Data analysis

In order to understand the data, we plot some visualizations before going through machine learning models. By counting the number of sentiments **(figure 2)** for each class we can conclude that tweets with negative sentiments are more frequent. Therefore, a first finding will be that Machine learning models that we will build will likely predict negative sentiments

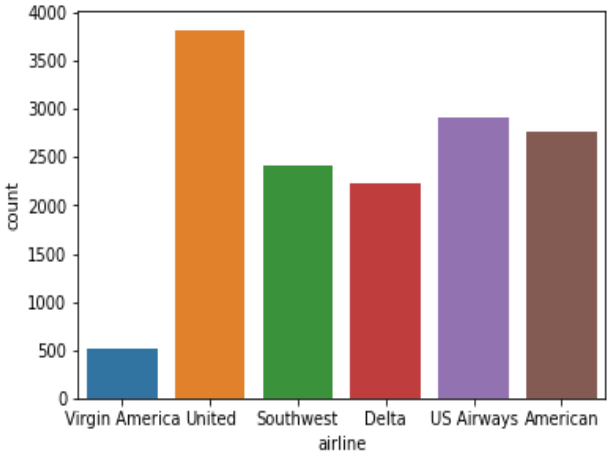


Figure 1: Number of tweets by Company

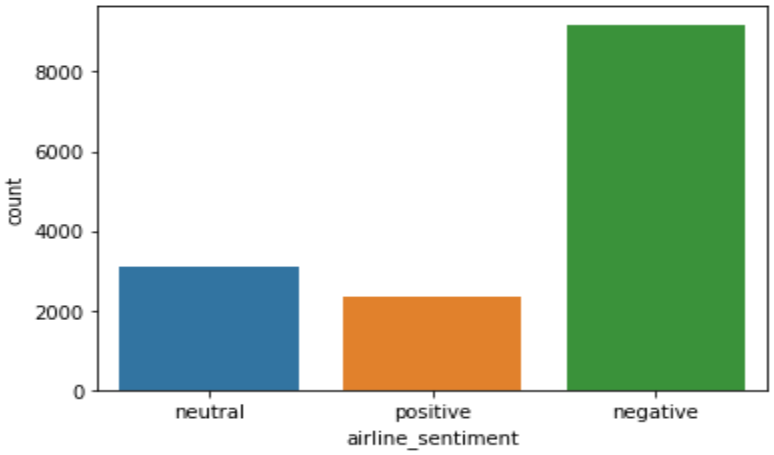


Figure 2: Number of tweets by sentiment

with more accuracy. It will be a good point because companies care more about negative reviews in order to improve their services. On the other hand, not all the companies have the same audience **(figure 1)** but as the customers express the same way we will not consider the company name in our analysis and we will focus on the content of tweets.

Data cleaning

Our approach is based on natural language processing where the input features are text tweets. Those tweets contain punctuation, lower cases, upper cases and stopwords which can affect the model learning performance. In brief the text cleaning involve:

- Removing punctuations : because it does not contribute to text analysis. Punctuation helps to make sentences readable but it affects the models' ability to differentiate between punctuation and other characters.
- All text in the tweets was converted to lowercase: Text analysis is case sensitive. So if we didn't change all text to lowercase "Good", and "good" would be considered as two different words for example.
- Removing stop words in the tweets. Stopwords have no analytic value for text analysis, Therefore they need to be removed to reduce the complexity.

Machine learning models

After data cleaning we train 3 machine learning models which are Random Forest, logistic regression and Adaptive boosting. Each model is trained and tested against 2 classification metrics which are recall and precision.

Results and Performances

We split the data to 75% train set and 25% test size. And we compute the precision, recall and we plot a confusion matrix for each model.

- Confusion Matrix: It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.

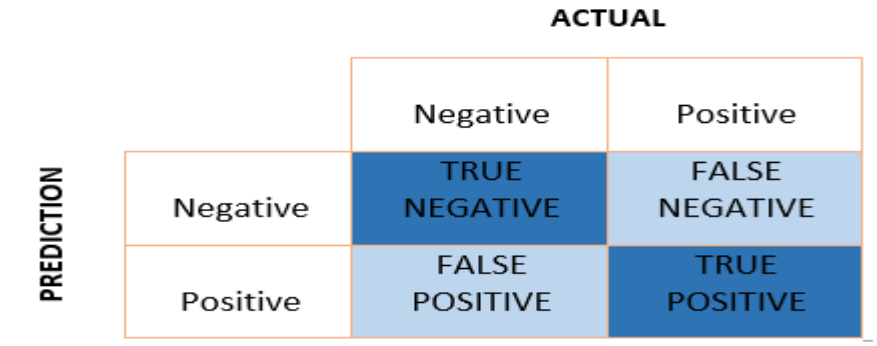
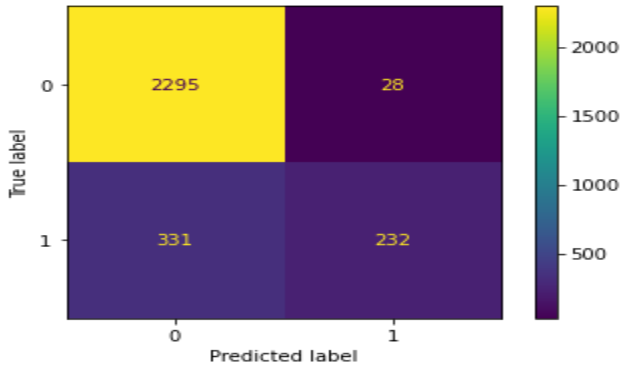


Figure3: Confusion matrix components

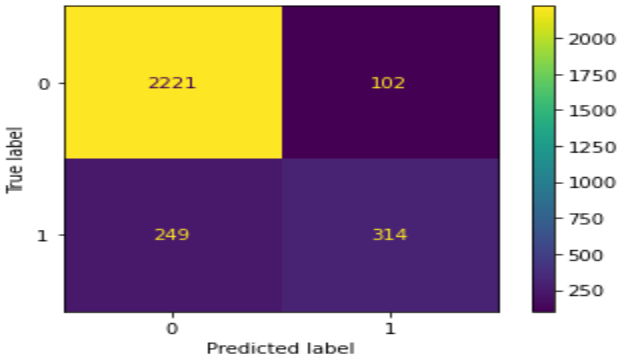
- **Precision** The ratio of correct positive predictions to the total predicted positives.
- **Recall** The ratio of correct positive predictions to the total positive examples.



Classification Report

	precision	recall	f1-score	support
0	0.96	0.90	0.93	2470
1	0.56	0.75	0.64	416
accuracy			0.88	2886
macro avg	0.76	0.83	0.78	2886
weighted avg	0.90	0.88	0.89	2886

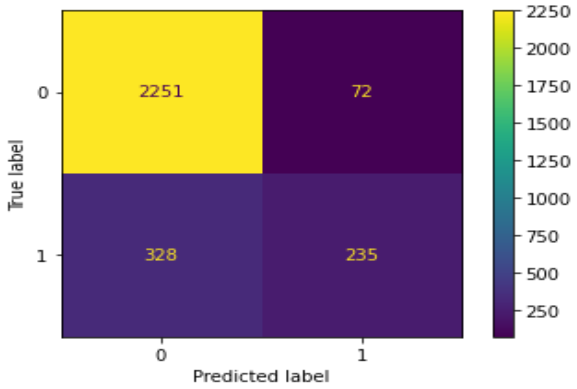
Figure 4: Results for logistic regression



Classification Report

	precision	recall	f1-score	support
0	0.99	0.87	0.93	2626
1	0.41	0.89	0.56	260
accuracy			0.88	2886
macro avg	0.70	0.88	0.75	2886
weighted avg	0.94	0.88	0.89	2886

Figure 5: Results for random forest



Classification Report

	precision	recall	f1-score	support
0	0.97	0.87	0.92	2579
1	0.42	0.77	0.54	307
accuracy			0.86	2886
macro avg	0.69	0.82	0.73	2886
weighted avg	0.91	0.86	0.88	2886

Figure 6: Results for Adaptive boosting

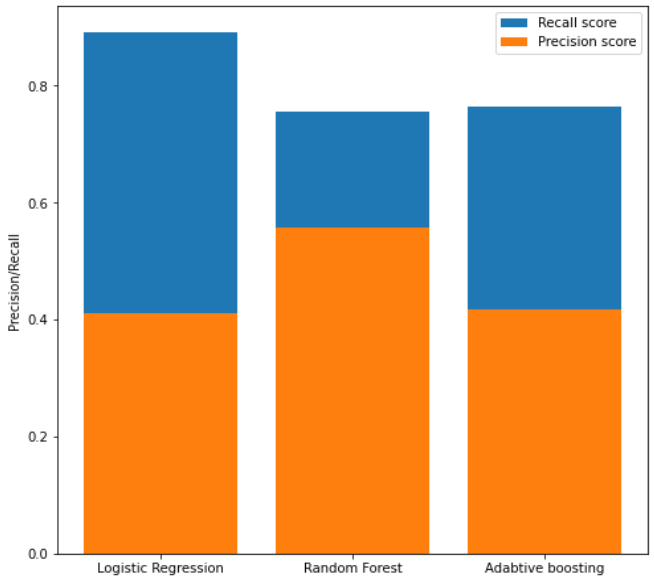


Figure 7: Comparison of 3 ML models.

We can observe that all the models have a high True negative rate which is good for our problem as we need to predict negative tweets more than positive tweets. **Figure 7** shows that Logistic regression has the highest Recall (which means a lower false negative which is the ideal case for our question).

Tools:

The tools used in this projects are:

- **Jupyter notebook :** It is an open-source web application that allows to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
- For data preparation we use the following python libraries : **Pandas** for data loading, **nltk** for text processing, **matplotlib** and **seaborn** for visualization.
- For machine learning models and performance evaluation we use **scikit-learn library.**