Airline Tweet Sentiment Analysis

Rika Yatchak

March 27, 2019

The problem:

Can we perform a sentiment analysis on customer tweets directed at selected U.S. Airlines?

Dataset source and description

The Twitter US Airline Sentiment was uploaded to Kaggle in 2016 by Figure Eight. It consists of two files:

1. Tweets.csv – Tweet text with tweet ID, date, timezone information, as well as labeled sentiments provided by Crowdflower.
2. database.sqlite – Same contents as Tweets.csv but in a sqlite database.

Data collected in 2015.

Methodology of data collection not clear: was the data cleaned? Were certain tweets removed?

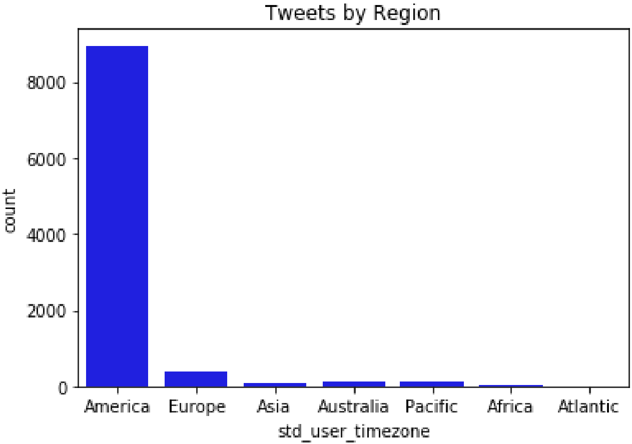
The Basics

14,467 tweets that @mention United, Virgin America, JetBlue, US Airways, American Air, Southwest Air, and Delta.

Corpus: 15,131 (including hashtags and mentions)

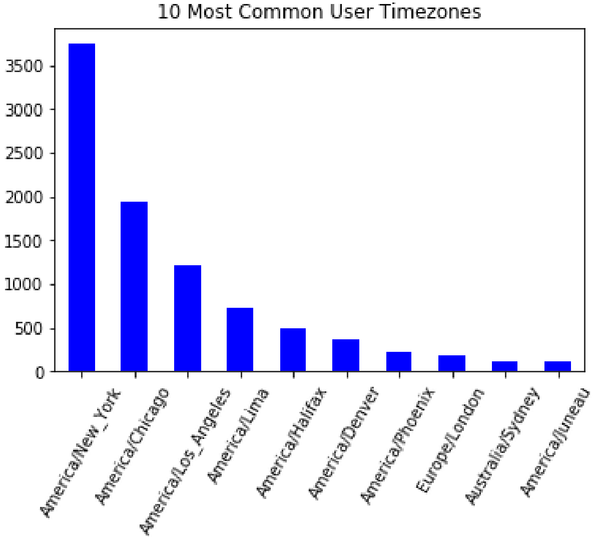
Tweets: average length is about 100 characters.

Where do tweets originate from?

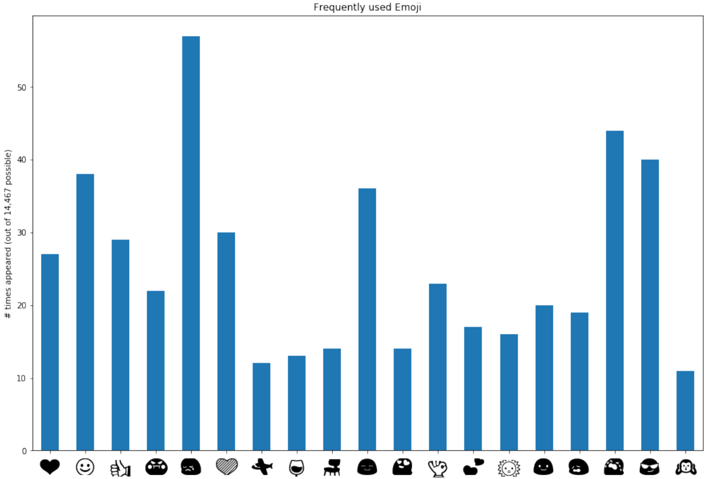


Most tweets originate from U.S. users.

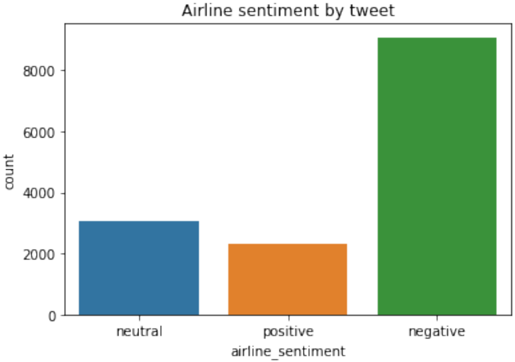
Most common Timezones



What kind of Emojis appear in this corpus?



What is the overall sentiment of the tweets?



Methodology

- Try three different kinds of inputs:
  - TF-IDF vectors
  - Word2Vec embeddings: sum of words and mean of words
  - Doc2Vec Paragraph (actually tweet) vectors
- Balance unbalanced sentiment classes using SMOTE: Synthetic Minority Oversampling Techniques. This technique simultaneously oversamples the minority classes while undersampling the majority class, which generally achieves better classifier performance.
- Try logistic regression and random forest classifier (scikit-learn implementations)

Results

Best performance: tuned random forest model with estimators and a maximum depth of 90.

Maximum test accuracy of .754 using a Sum of Words embedding.