Twitter Airline Sentiment Analysis

Haonan Zhang
8th June 2020

Motivation

By Analyzing the twitter airline reviews:

1. Help airline companies to predict whether passengers have positive attitude or negative attitude about the flights and services

2. Identify potential business opportunities, risky areas, and other business opportunities to improve the airlines' economic performance and passenger loyalty

Demo

Data

Scrapped by a company called Crowdflower, and publicly available at Kaggle

Human labels the sentiment categories

MySQL in RDS is used for storing predictions whenever a entry is submitted on website.

A tweet_sentiment_prediction table is built.

Feature Engineering

Tokenization: separate texts into tokens for a bag of words representation

Remove uninformative punctuation and stopwords

Stemming and lemmatization

Bag-of-words model(based on word counts)

Term frequency Inverse Document Frequency(TF-IDF)

Model - Logistic Regression

Model the log odd of being positive sentiment over negative sentiment by a linear combination of text features

Create 1520 text features via training data using tf-idf vectorizer or bag-of-words vectorize from scikit-learn package

Using logit function transformation to obtain the probability of being positive sentiment via a linear model.

Train Test Split: 70%/30%

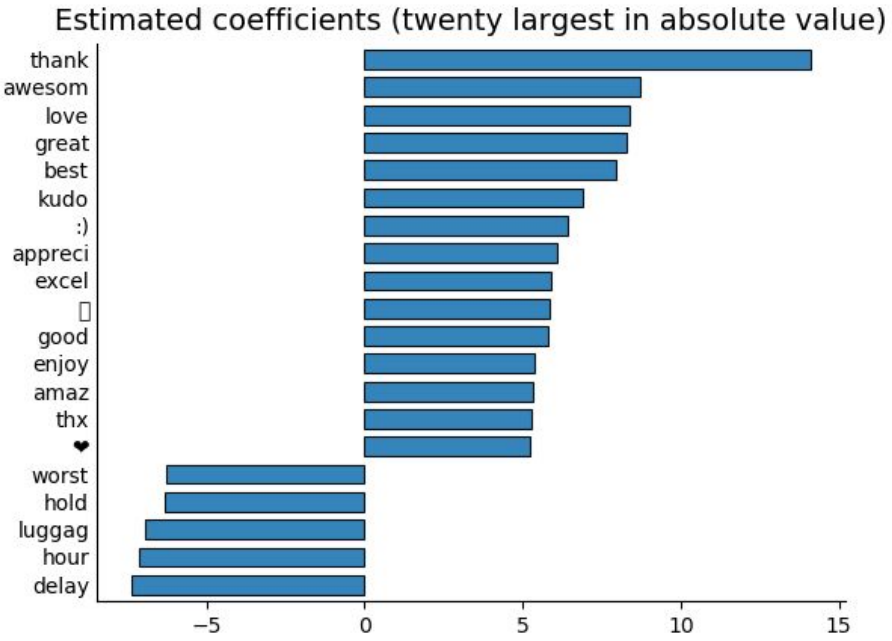
Success Criteria AUC > 0.8

Test AUC: 0.975

Test Accuracy: 0.944

F1-score: 0.810

Insights



The model provides a sufficiently good test accuracy.

The model is interpretable, such as correctly identifying positive and negative words.

Other advanced tools can help extract entities related to service and customer service.

Thank you