Capstone 2: Milestone Report

Rika Yatchak

March 26, 2019

0 Introduction

In this project, I examine the Twitter US Airline Sentiment for my Springboard Capstone project. To view the code used to generate the results in this report, see the corresponding Git repository.

In this report, we will

1. Define a guiding question for the project.
2. Identify hypothetical clients.
3. Describe the dataset and the cleaning/wrangling steps needed to prepare for analysis.
4. List other potential datasets that could be used.
5. Explain the findings.

1 Project Question

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

2 Hypothetical Client

The hypothetical client is a an airline which would like to support its social media managers in quickly identifying tweets that should be prioritized for immediate response. They would like to identify upset customers who may need additional support using an automated system that flags tweets that are negative and directs customer service representatives to review them for possible issue resolution.

3 Dataset Description and Preparation

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

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

The dataset appears to have been collected sometime in 2015. Unfortunately, the methodology for data collection and postprocessing was not published with the dataset. From my EDA, it seems clear that the dataset was collected by looking for tweets that @ mention the official accounts of various U.S. airlines. There are some columns that were added after data collection: the columns `airline_sentiment` and `negative_reason` appear to be the output of Crowdflower's own sentiment analysis and topic categorization.

As this dataset contains tweet text, significant postprocessing was needed: emojis were identified and removed, various regex expressions were utilized to remove undesired parts of speech or text snippets such as URLs and personal email addresses, and hashtags were also post-processed. The data cleaning notebook has more details about the steps that were undertaken to prepare text for sentiment analysis.

4 Other Datasets

Since the main bulk of the data is output from the Twitter API, there is always the possibility to obtain more relevant tweets from Twitter itself. To obtain more data, we would need to pull tweets that @ mention Virgin America, United, US Airways, JetBlue, and Southwest Air. Since the dataset was collected, Virgin America has ceased operation. We could also consider adding airlines such as Delta, Frontier, and Alaska Airlines, which are not included in the current dataset.

The main reason that more tweets were not pulled for this project is the dramatic increase in data wrangling and data cleaning that this would require: the tweets made available by Crowdflower in this dataset appear to have been selected especially for sentiment analysis and have been partially cleaned.

5 Findings

We found that the tweets contained in this dataset tend to be negative: roughly 62% of the tweets were found to be negative by Crowdflower. See the Data Story for more details about the dataset. The most-used emojis, hashtags, and @mentions in the dataset are consistent with tweets from customers to airline official accounts: most tweets are overwhelmingly positive or negative.