Airline Tweet NLP

Jai Bansal

May 18, 2016

This document conducts basic natural language processing (NLP) analysis on Twitter tweets about major US airlines scraped from the site during part of February 2015.  
Contributors were asked to classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). Thus, for each tweet, I have the 'correct' answer for sentiment analysis purposes.

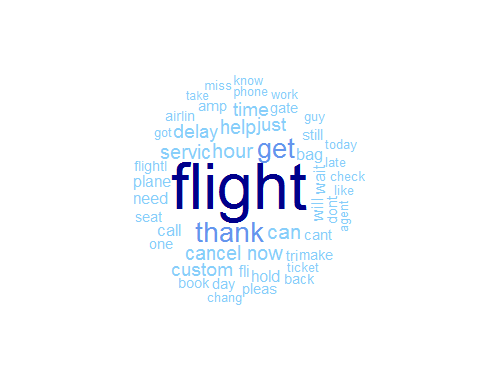
The data can be found at the URL below. To find the dataset, search for 'Airline' on the page.  
I specifically use the 16,000 row dataset uploaded on February 12, 2015 by CrowdFlower.  
I assume the upload date is incorrect as the data includes tweets from after 2/12/2015...

<https://www.crowdflower.com/data-for-everyone/>

Note that the actual dataset only appears to contain 14,640 rows. I'm not sure where the discrepancy between actual rows and rows stated on the website comes from, but it doesn't affect my analysis.

### Word Cloud:

Below is a word cloud with the 50 most frequently used words (technically stems) in the tweet data. The larger and darker a word, the more frequently it was used.



### Sentiment Analysis:

I will conduct sentiment analysis in 2 ways.  
\* Lexicon based (with pre-provided lists of positive and negative terms)  
\* Naive Bayes Classification Model

**Lexicon Based**:

In this section, I use a function that automatically preprocesses the data and computes various scoring metrics.  
The preprocessing that is done includes making the text lowercase, removing punctuation, removing numbers, removing (English) stopwords, removing whitespace, stemming words, and removing words with less than 3 characters.  
The scoring metric I will consider is polarity and is computed as: (p - n) / (p + n)  
p[n] is the number of positive[negative] words in a tweet. These positive and negative words come in a pre-defined list.

For each tweet, if polarity is less[greater] than 0, the tweet will be predicted to have negative[positive] sentiment.  
Tweets will a polarity of 0 will be predicted to be neutral.

Using these definitions, 5544 out of 14640 (37.87)% of tweets are predicted correctly.

Broken down by category:  
1350 out of 2363 (57.13)% of positive tweets are predicted correctly.

1585 out of 3099 (51.15)% of neutral tweets are predicted correctly.

2609 out of 9178 (28.43)% of negative tweets are predicted correctly.

Further Exploration:

These results are not great. Results might be improved by using a different (possibly custom) list of positively and negatively associated words. Since these tweets are directed at airlines, the lists should probably include air travel specific terms. Results might also improve by using different polarity cutoffs. Perhaps tweets with scores that are slightly above/below zero should be classifed as neutral. Or maybe tweets with a slightly negative polarity should be classified as positive. There are many variations to explore here. Finally, a different scoring metric might yield better results.

**Naive Bayes Classification Model**:

The training error is 0.5855191.  
The test error is 0.5840164.