Airline Tweets NLP

Jai Bansal

May 24, 2016

This document shows the results of basic natural language processing (NLP) analysis on Twitter tweets about major US airlines scraped from the site during part of February 2015. Specifically, I create a word cloud and conduct sentiment analysis.

Contributors to the data set were asked to classify positive, negative, and neutral tweets. Thus, for each tweet, I have the 'correct' answer for sentiment analysis purposes.

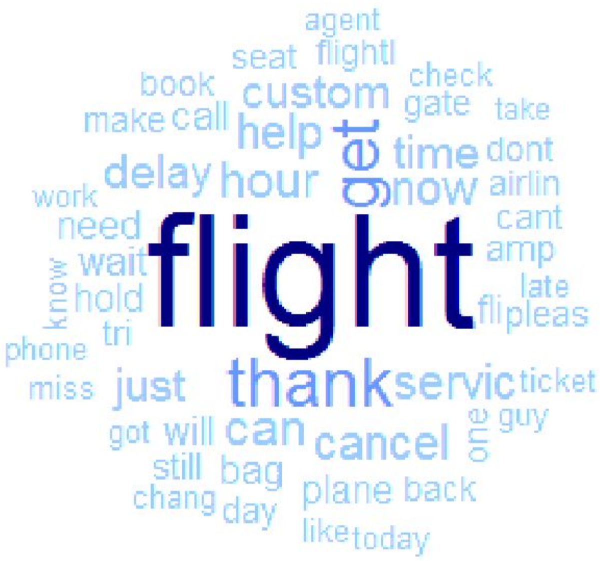
The data can be found at the URL below. To find the dataset, search for 'Airline' on the page. I specifically use the 16,000 row dataset uploaded on February 12, 2015 by CrowdFlower. I assume the upload date is incorrect as the data includes tweets from after 2/12/2015...

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

Note that the actual dataset only contains 14,640 rows. I'm not sure where the discrepancy comes from, but it doesn't affect the 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 conduct sentiment analysis in 2 ways:
* Lexicon based (with pre-provided lists of positive and negative terms)
* Naive Bayes Classification Model

Lexicon Based:

Before modeling, I make the text lowercase, remove punctuation, remove numbers, remove (English) stopwords, remove whitespace, stem words, and remove words with less than 3 characters.
The scoring metric I use 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 classified as negative[positive].
Tweets with a polarity of 0 are classified as neutral.

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

Results by tweet classification category:
1350 out of 2363 (57.13)% of positive tweets are classified correctly.

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

2609 out of 9178 (28.43)% of negative tweets are classified 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 classified as neutral. There are many variations to explore here. Finally, a different scoring metric might yield better results.

Naive Bayes Classification Model:

After computing feature importance for each word, I include the 110 words with a non-zero (actually, greater than 0) feature importance.

Results:
Training error: 58.55%
Cross Validation error: 58.61%
Test error: 58.4%

Results by tweet classification category:
1086 out of 1192 (91.11)% of positive tweets are classified correctly.

252 out of 1516 (16.62)% of neutral tweets are classified correctly.

1707 out of 4612 (37.01)% of negative tweets are classified correctly.

Overall, the Naive Bayes model performs slightly better than the lexicon based model. Interestingly, the Naive Bayes model does extremely well on positive tweets and much better than the lexicon based model. However, Naive Bayes does poorly on neutral tweets and much worse than the lexicon based model. Naive Bayes does somewhat better than the lexicon based model when classifying negative tweets.

Interestingly, training, cross validation, and test set error are almost identical.

Further Exploration:

Better results might be achieved by using a different feature engineering process (more or less features) or by tuning some Naive Bayes model parameters (the laplace smoothing parameter is one option).