Airline Sentiment Analysis

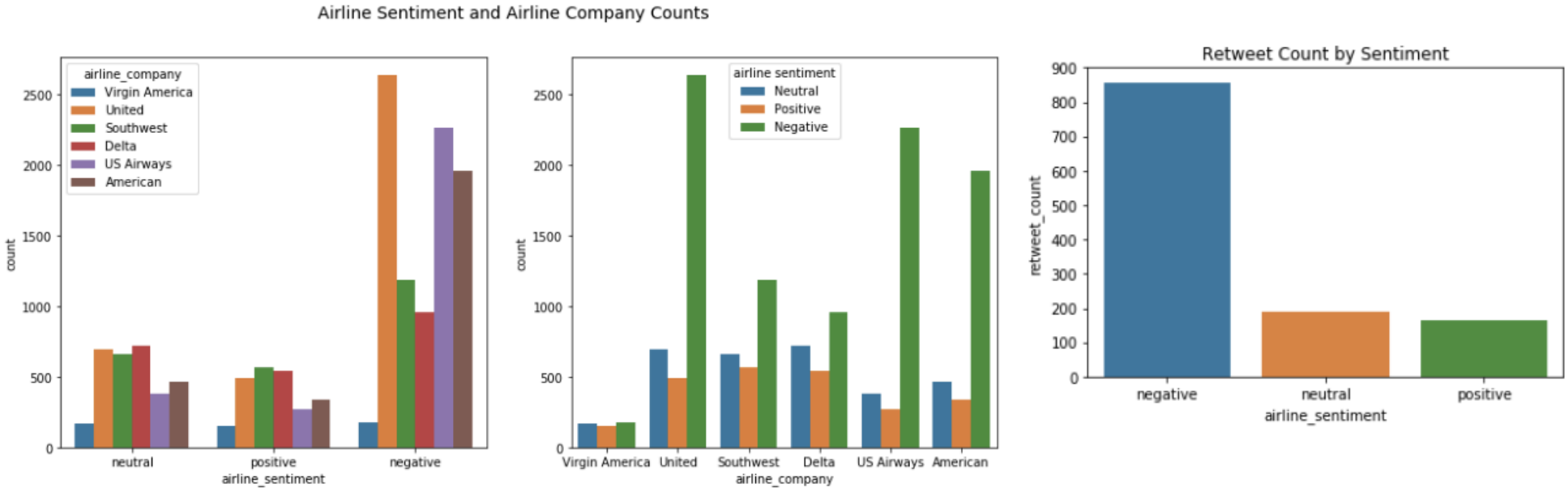
Bo Liu
Springboard Data Science
March 2020

Table of Contents

Problem Statement & Data Acquisition

- Social media such as Facebook and Twitter play an important role in people's lives today. Users share product and service reviews on social media.
- This project focuses on sentiment analysis on a few major U.S. airlines. The goal is to predict whether a customer review is neutral, positive or negative.
- Given the customer reviews and sentiment labels provided by the dataset, supervised classification models are built to predict the probability of reviews falling into 3 classes of sentiments.
- Data source: https://www.figure-eight.com/data-for-everyone/
- Raw Data: A csv file with customer reviews, sentiment labels, number of retweet counts, airline company, etc. Total number of reviews is ~15,000.

Data Visualization



- A visual inspection shows that majority of the sentiments are negative.
- There are significantly more retweets in the negative sentiment group than the other 2 sentiment groups. Including the retweet column may add value in the classifications.

Text Cleaning

Sentiment	*Text*	*Airline*
neutral	@VirginAmerica I didn't today... Must mean I need to take another trip!	Virgin America
negative	@VirginAmerica it's really aggressive to blast **obnoxious** "entertainment" in your guests' faces & they have little recourse	Virgin America
negative	@VirginAmerica and it's a really big **bad thing** about it	Virgin America
negative	@VirginAmerica **seriously would** pay $30 a flight for seats that didn't have this playing.	Virgin America
positive	@VirginAmerica yes, nearly every time I fly VX this ‰ьIIear worm‰ьIќ won‰ьIЄt go away **:)**	Virgin America
neutral	@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP	Virgin America
positive	@virginamerica Well, I didn't‰ьI_but NOW I DO! **:-D**	Virgin America

Text
today must mean need take another trip
really aggressive blast obnoxious entertainment guest face little recourse
really big bad thing
seriously would pay flight seat playing
yes nearly every time fly vx ear worm win go away
really miss prime opportunity man without hat parody
well

Before

After

Steps of text cleaning:

- Removing hashtags and mentions
- Lowercase all words
- Removing punctuations, stop words and non-English words
- Removing URLs in the text
- Expanding contractions
- Lemmatization
- Remove NaN and blank text fields

Modeling

Feature Extraction + Model	AVG Accuracy (STD)
BOW + Random Forest	0.638 (0.006)
BOW + Naïve Bayes	0.760 (0.009)
BOW + LightGBM	0.764 (0.004)
BOW + Logistic	**0.788 (0.006)**
TF-IDF + Random Forest	0.633 (0.007)
TF-IDF + Naïve Bayes	0.702 (0.006)
TF-IDF + LightGBM	0.761 (0.002)
TF-IDF + Logistic	0.764 (0.003)
BOW + 2-Gram + Naïve Bayes	0.747 (0.006)
Word2Vec + Random Forest	0.706 (0.006)
Word2Vec + LightGBM	0.708 (0.003)
Word2Vec + Logistic	0.729 (0.007)

NLP feature extraction techniques:

- CountVectorizer (BOW)
- BOW + N-Grams
- TF-IDF Vectorizer
- Word2Vec

Model within each framework:

- Random Forest
- Multinomial Naïve Bayes
- LightGBM
- Logistic regression

For each feature model combination:

- 5-fold cross validation using RandomizedSearchCV
- Average accuracy scores on the validation sets are recorded and compared among different feature model combinations.
- Best combination is CountVectorizer + Logistic Regression

Model Evaluation

	precision	recall	f1-score	support
0	0.59	0.53	0.56	615
1	0.76	0.62	0.68	485
2	0.82	0.89	0.85	1811
micro avg	0.77	0.77	0.77	2911
macro avg	0.72	0.68	0.70	2911
weighted avg	0.76	0.77	0.76	2911

- CountVectorizer + Logistic Regression combination has the highest average accuracy score on the validation dataset, and hence selected to be the final feature + model recommendation.
- We evaluated this model on the independent testing set, which is a random 20% held out of the original dataset. Overall accuracy on the test set is 0.7695.
- Using 0.5 as cutoff threshold, precision and recall for each target group are shown in the report above. Note that 0 represents neutral sentiment, 1 meaning positive and 2 meaning negative.

Model Evaluation

Examples: False positive (actual labels are negative)

@JetBlue the free wifi makes up for the television not working... It's staticy #ithelpsabit
@united thanks for the link, now finally arrived in Brussels, 9 h after schedule...
@united thank you for dishonoring my upgrade and putting me in a seat I didn't want, all while not even notifying me. Great 1K service ЩΦ
@USAirways Forget reservations. Thank you to the great leadership at your company, I've Cancelled Flighted my flight. Once again, thank you.

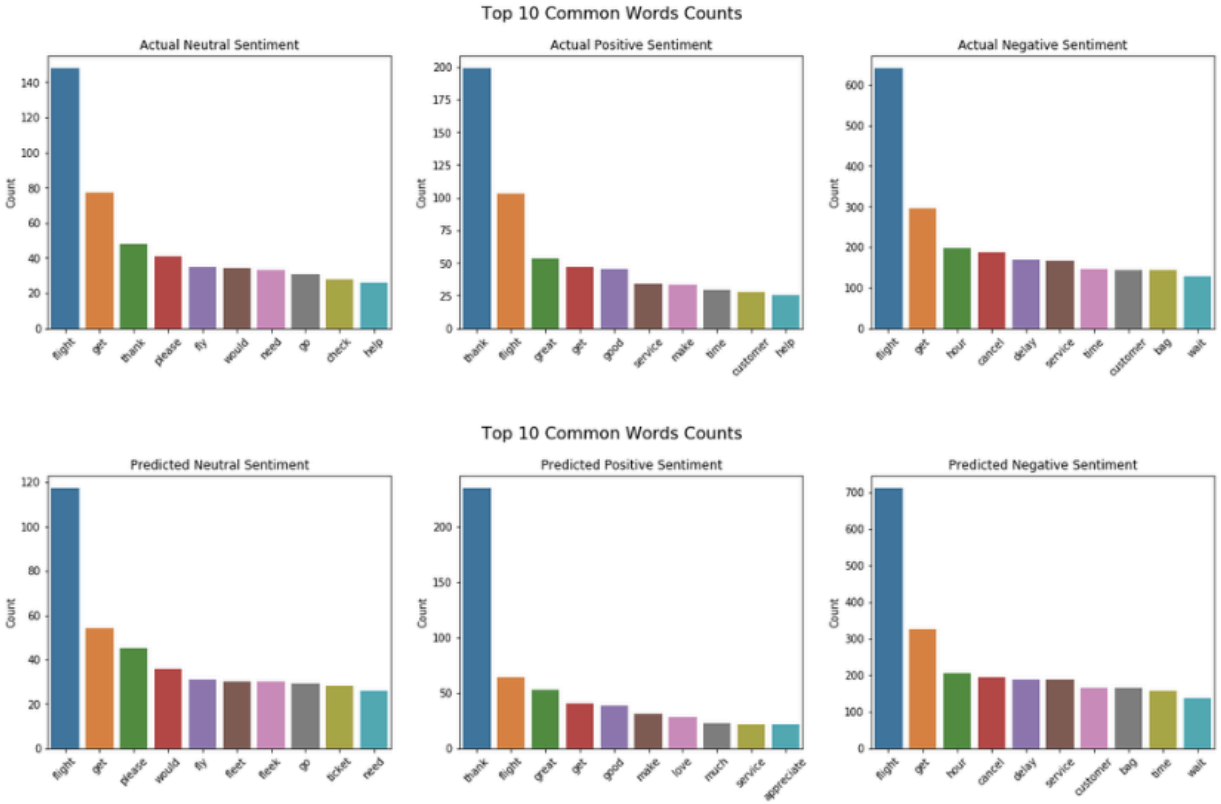
- Many have the word “thank” but actually expressing dissatisfaction of the service. Interestingly, we believe the last review in these examples should be a positive one just as our model had predicted, but the actual label provided in the dataset is negative.

- In the first review below, the customer first expressed his/her thanks to 2 airline associates at baggage claim, but then said he/she has a complaint. The model captures the negative sentiment.
- in the last review the customer took a United flight and says united suck, so this should be a negative sentiment, rather than positive. Hence, we believe the original label is wrong.

Examples: False negative (actual labels are positive)

@USAirways please thank Mellie at CAE, Tammy in baggage claim at CLT 4 #excellent customer service 2day, BUT I have a complaint.
@JetBlue Touchdown JFK! Well done pilots of JetBlue Flight 226! #JetBlueRocks
@united Brian at SFO customer service deserves a raise, gave me extra meal voucher and a good joke to cheer me up after flight delay. #FTW
@united you suck. @SouthwestAir you're the best.

Model Evaluation



- Graph shows top 10 common words in the actual and predicted sentiments.
- Predictions are pretty close to the actual sentiments. For example, there are 7 words overlapping in the actual and predicted positive sentiment common words, and we can see words such as “thank”, “good”, “appreciate” that best captures positive sentiments are selected by the model.

Summary

Data & Modeling:

- Data pulled from figure-eight website where ~15,000 Twitter reviews of 6 major American Airline companies are scraped. Reviews are from Feb 2015 time frame.
- EDA and text cleaning are performed to format the features into table or matrix that machine learning algorithms can process. Steps include removing hashtags and mentions, removing URL links in the review, removing punctuations, stop words and non-English words, expanding contractions, lemmatization, removing NaN and blank text fields.
- Several combinations of NLP feature extraction methodology and classification models are compared on the validation datasets and the combination with the highest accuracy on the validation data is CountVectorizer + Logistic Regression.
- We evaluated the best model on the independent testing set, which is a random 20% held out of the original dataset. Overall accuracy on the test set is 0.7695.

Business implications:

- For each Twitter review, the model outputs probability of neutral, positive or negative sentiment. For better customer experience, analysts from airline companies can collect the predicted negative reviews and proactively reach out to a selected group of customers (eg. VIP) to resolve the problems. This will eventually reduce customer turn over rate.

Future Work

- Explore emoticons in data processing and modeling.
- There are many mis-spellings in the documents, explore ways to correct mis-spellings.
- Try a few more combinations of hyperparameters in the Word2Vec model in Gensim, such as tweaking window size and total number of features to keep. It is also a good idea to train Word2Vec model using the airline dataset using neural network (NN), rather than using pre-trained models in Gensim.
- Building more sophisticated NN models such as Recurrent NN using more curated feature matrices.
- The model often makes predictions based on a single word in the sentence that it recognizes as positive or negative, but lacks overall understanding of the sentences like what humans can do. In a future refinement, we can train the model to understand the contexts of the reviews by using some topic modelling like Latent Dirichlet Allocation or Probabilistic Latent Semantic Analysis.

Acknowledgement

- Thyago Porpino (mentorship)
- Kaggle (open data source)
- Springboard team (curriculum & administrative support)

Reference

Data Source:
https://www.figure-eight.com/data-for-everyone/
Data Wrangling & EDA Notebook:
https://github.com/lisalb168/Airline_Sentiment_Classification/tree/master/notebook
Final Report:
https://github.com/lisalb168/Airline_Sentiment_Classification/tree/master/report