Project 4 Summary

Gaining Insight on Customer Topics of a US Airline via Twitter

(Natural Language Processing)

Charlie Lew

November 18th, 2019

Abstract

This project uses Natural Language Processing (NLP) to analyze customer topics from a series of tweets collected from Twitter for Delta Airlines. It is no secret that North American airline companies rely heavily on social media to not only address customer concerns but also advertise current and upcoming products and services. The collected tweets span over two weeks from late October to mid-November 2019. The pipeline process was collect, clean, vectorize and then topic-model. Several interesting observations were noticed. The most common one was complaint about delays. Future work on this project involves further cleaning, gathering data, real-time topic and sentiment analysis and deep learning to filter out the noise.

Introduction

In 2017, the airline response rate to customer queries on social media was 21%. In 2018, that number rose to 24%. Hence, it is no secret that domestic airlines are relying heavily on social media not to only address immediate customer concerns but also use it as a tool to offer new product and services. The latter being the current interest of this study. Can real-time social media feeds be used to gather information on customer chatter to decide on what services to offer them in the short and long term? The general consensus is to immediately address a customer complaint. Hence, this is the ultimate goal of this study by first analyzing topics from chatter on Twitter and then utilizing it later on for marketing purposes. The following summary is then divided as follows: the next section discusses the tools, data and methodology. The results section discusses the findings and it finally closes with a future work.

Data, Tools & Methodology

For this case study, Delta Airlines was picked for two main reasons. First, it has quite a large market share of nearly 20% in North America. A close second is American Airlines which could also have been picked but for the purposes of this study it may not necessarily matter. Second, its daily twitter volume averages 1,500 not counting re-tweets. Tweepy which is a python Application Programming Interface (API) was used to download fresh tweets from Twitter using the handle of @Delta. The number of tweets totaled nearly 23,000 spanning

dates between October 29th to November 13th, 2019. Figure 1 below shows the methodology pipeline and tools more clearly.

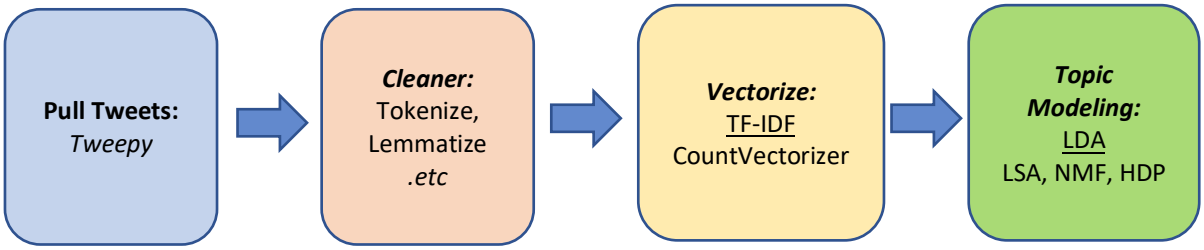

Figure 1: Methodology pipeline of tools for NLP tweet analysis

As mentioned, the tweets are first gathered from Twitter via Tweepy. The collected tweets are then 'cleaned' whereby, stop words, Byte Order Marks (BOM), numbers, full stops, hash tags (#), alias handles (@), exclamations points, etc are removed. Second, after those are processed, the texts are then lemmatized, and tokenized for the next step of vectorizing. Here, the Term Frequency-Inverse Document Frequency (TF-IDF) methodology was used as opposed to the CountVectorizer as it gave the best overall result when paired with a topic modeler. Hence, the underlined tool of TF-IDF in Figure 1. Dimensionality reduction was achieved via Latent Dirichlet Allocation (LDA) although Latent Semantic Analysis (LSA) which is Singular Value Decomposition (SVD), Non-Negative Matrix Factorization (NMF) and Hierarchical Dirichlet Process (HDP) were used but not chosen as LDA with TF-IDF provided the best subjective topic result. As a note, there no clustering was performed.

Results

I arrived at 20 topics from LDA but chose six that were interpretable. Figure 2 below shows a bar chart arranged by the number of topics from highest to lowest.

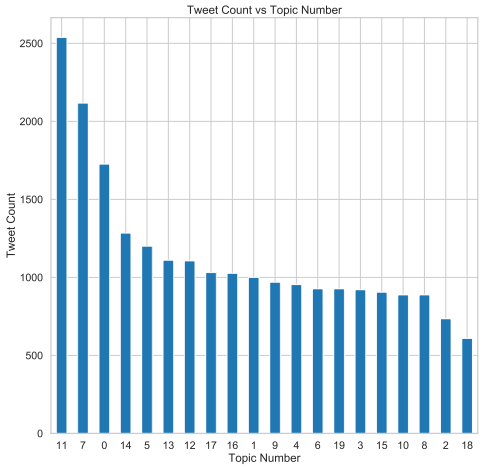

Figure 2: Tweet count vs Topic number

From Figure 2, Topic 11 has the highest relatable number of tweets with a little over 2,500 which is more than 10% of the tweet population. Table 1 below shows four tweets selected from the pool.

Table 1: Topic and top five words (10/29 to 11/13)

Topic Number	Word 1	Word 2	Word 3	Word 4	Word 5
11	Hour	Delay	Minute	Connection	Waiting
17	Baby	Pump	Breast	Bathroom	Milk
1	Porn	Child	Movie	Watch	Showing
2	Hope	Shame	Damn	Breed	Service

Topic 11 pertains mostly to chatter about customer complaints for flights being delayed and missing their respective connections. In fact, the most complaints any airline in the world will receive are delays. Hence, Delta is no stranger to delays though any airline does the best it can to avoid this. Topic number 17 pertains to customers complaints about not being able to find proper place within airports to properly breast pump or breast feed their babies. The other side of the argument is that customers are complaining about seeing mothers breast feed in the open with no cover. Hence, with a complaint like this, Delta could lobby airport authority to erect locations for mothers to breast feed their babies. Topic 1 on the other hand was an incident whereby the movie selection in the airplane had pornographic content in it and a few children were able to access them making parents furious at Delta. Instances like these can be avoided if airlines properly put in place a rigorous screening process for movies or completely remove pornographic content on their airplanes in the future. Finally, Topic 2 discusses customers who were unaware that not all breeds of dogs are allowed as service dogs on-board. One instance was a pit bull. Delta and the FAA has to be really clear on what type of service dogs are allowed and must let the passenger know in advance what the restrictions are.

Closing Remarks & Future Work

Natural Language Processing and in particular Topic Modeling is a challenging task. Cleaning text could potentially change the topic outcome and interpretability. Tweets can be inherently noisy in sense that a lot of chatter is not necessarily directed at Delta per se but users using Delta's tweet medium as a talking platform. Nonetheless, the topics discussed in the above summary properly captured the preliminary sentiment of the last two weeks. Future work include cleaning the text and obtaining more data. As time progresses so the type of topics. An airline could potentially benefit by processing real-time tweet data and providing these topics along with sentiment to marketing for future offerings to their customers.