Airline Customer Sentiment Analysis using Tweets

IS 8070 – Survey of Machine Learning

Submitted by
Hridhay Mehta
Priya Kumari
Utkarsh Singh

Contents

At a Glance

Increasing turbulence in the Airlines Industry and a competitive landscape are pushing all operators to rethink customer strategies. A key enabler is the data available on customers.

Importance of customer voice on social media

Travelers nowadays are increasingly turning to social media, primarily Twitter and Facebook, to express their travel experience. Once posted on social media, a matter is no more limited to the particular consumer and service provider only, it is broadcasted to a much larger audience. It also becomes a source for other companies to get insights into their competitor's business.

Need for social media sentiment analysis

Sharing opinions on social media has put an additional time pressure on airline companies to detect and mitigate issues before they become a PR nightmare. A single mishandled incident can get viral on social media in no time. Therefore it is important to predict the potential risks and changing customer sentiments proactively.

Milestones for a social media sentiment analysis project

For a project aiming to understand consumer sentiments on Social Media, the first milestone is to collect natural language data. Project teams also need to generate a "response" metric depending on the objective – i.e. a way to tell the machine what is good or bad. The second milestone is to develop models that can accurately predict the response based on the available data. The third milestone is a strategy to translate model predictions into a set of actions needs to be put in place. This will include what needs to be done and who will do it (Eg. Dedicated social media management teams). The fourth milestone is to develop a set of KPIs that will help monitor the effectiveness of strategic initiatives.

Social Media Landscape

Over the past decade, consumers have largely shifted from scanning traditional print and TV media to using Internet and social media sites like Twitter for current affairs. Following this trend, even the mainstream journalism and media houses have increased focus on social media subscriptions. As a result, social media has become a popular platform for consumers to exchange knowledge about their individual experiences.

At present, Twitter and Facebook are the two most prominent social media platforms for text-based user activity. Social media adoption is >80% across the four generations that live today in the US[1]. Worldwide, there are over 1.66 billion daily active users on Facebook and 145 million daily active users on Twitter.

Although Facebook currently has a much larger active user base than Twitter, there are significant differences between the purpose for which both channels are used. Also, Twitter has become increasingly popular for live events, public updates and news. Facebook encourages its users to personalize content and accordingly make it accessible to a relatively smaller network, however Twitter is much less private and makes it easier to expand network. Journalists, Politicians, Senior Executives of Companies and Mass influencers are more active on Twitter. About 74% of Twitter users say they use Twitter to get their News[2].

There is also a significant difference between the adoption of these platforms across different user demographics. Twitter's popularity has increased amongst the younger user group.

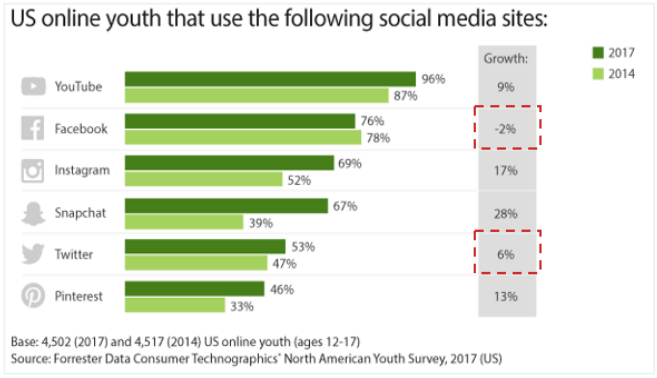


Figure 1: 6% icnrease in Twitter adoption vs 2% reduction in Facebook adoption

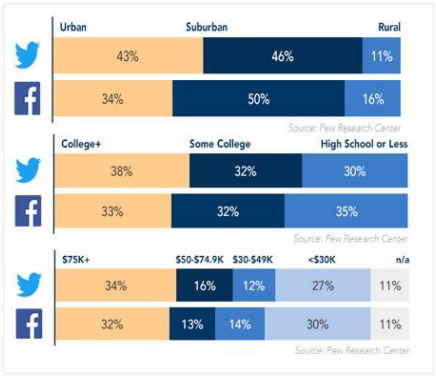


Figure 2: Demographics comparison

A higher proportion of Twitter users are more educated and earn higher than the Facebook user base. Also, about 47% of Twitter users dwell in Urban areas as compared to 35% for Facebook users.

Social media broadcasting can leave a powerful impact on the brand of organizations that provide consumer services of any form. It is also the most efficient way of reaching out to larger audience in a short timeframe. Therefore, organizations too have been maximizing social media adoption through their Facebook pages or Twitter handles

1 - https://www.business2community.com/social-media/how-the-social-media-landscape-is-changing-in-2019-02242391
2 - https://blog.hubspot.com/marketing/twitter-vs-facebook

Relevance for Airlines Industry

On 10th April 2017, the United Airlines staff dragged a customer out of an overcrowded airplane. A live recording of the incident went viral on social media, and the internet was flooded with hostile customer sentiments within hours. One of the videos received over 7 million views within 24 hours. Over the next 7 days, the company's stock price dipped by 6%[3]. The CEO issued a public apology on Twitter. Over 143,000 people reacted to the post and over 90% with an angry emoticon. On Twitter, the United page saw 20000 more followers over 10 days, which usually takes about 2 months. This example illustrates how important it is for the Airlines industry to monitor social media sentiments.

The Airlines Industry is very active on social media. According to ***The Atlantic,*** JetBlue connects with more than 20 people per hour. American Airlines receives more than 4,500 mentions per hour – 70% to 80% on Twitter. In order to drive positive customer sentiments, Airline marketers are increasingly generating content on social media. In 2016, the Airlines Industry was publishing on average 6-7 pieces of content (Updates, upcoming offers or ad campaigns) every month. In June 2017, KLM airlines started offering flight data over Twitter. Passengers can request booking confirmation, view check-in information, get a boarding pass, and view flight status on social media platforms.

When a customer tweets about an airline operator and receives a response, they are willing to spend on average $9[5] more on average-priced services from that operator in the future. This increases to $20 if the response is within 6 minutes[4].

Figure 3: Increase in revenue potential per transaction - Airlines Industry

3 - https://www.macrotrends.net/stocks/charts/UAL/united-airlines-holdings-inc/stock-price-history
4 - https://www.forbes.com/sites/mckinsey/2015/07/01/social-care-in-the-world-of-now/#4dd2affb35a8
5 - https://blog.twitter.com/en_us/topics/insights/2016/study-twitter-customer-care-increases-willingness-to-pay-across-industries.html

What is Sentiment Analysis?

Sentiment Analysis is the automated process of identifying and extracting the subjective information that lies in a text. The most common type of sentiment analysis is called "Polarity detection", which results in classifying a piece of text as 'positive', 'negative' or 'neutral'.

Sentiment Analysis finds application in the following fields:

- Social Media Monitoring
  - Prioritize actions to address raging issues
  - Track trends over time
  - Keep an eye on competition

- Brand Monitoring
  - Understand how brand perception evolves over time
  - Understand public reaction to brand awareness campaigns

- Voice of Customer
  - Understand nuances of customer experience through online surveys
  - Design better informed questions to ask on future surveys
  - Respond more quickly to signals from customers

- Customer Service
  - Prioritize order for responding to service tickets
  - Efficiently escalate in-progress cases to relevant service teams

- Market research
  - Tap into new sources of real-time information
  - Quantify otherwise qualitative information
  - Reduce expense of collecting structured data wherever possible
  - Fill in gaps where public data is scarce – eg. Customer's likes dislikes about restaurants, hotels etc.

Machine learning for sentiment analysis

Nearly 80% of the world's data is unstructured and this proportion would be higher for data scraped from social media. Since the information is not organized in a pre-defined way, it is difficult to sort and analyze data. Machine Learning has enabled users to create models that can learn on manually-labelled data and can be used to process unstructured data in future.

Text Mining techniques such as Bag-Of-Words (Count Vectorization, TF-IDF vectorization) or Word Embeddings can help extract features from processed text data. Supervised Learning techniques such as Random Forest, Boosting or Deep Learning can utilize these features to learn to predict a response.

However, building these models would require mapping a pre-determined sentiment label to each text input based on the context of the problem. This would be used as the "response variable" for training the Supervised Learning models.

Below are some possible ways by which airline operators can collect or create a sentiment response variable for the set of corpuses they wish to input in the model:

1. Manually scan through texts and tag a response using manual judgement. This can be a time-intensive exercise and judgements involve subjectivity
2. Push an online survey requesting a sample of users to specify their sentiment (Happy/Neutral/Disappointed) if they have made a post. User response can then be mapped to the corresponding text
3. Web apps that enable customers to post about their experiences real-time along with an overall response. This would help in collecting natural language data as well a response at the same time

A number of open source libraries support text mining for Sentiment Analysis based on lexicology and do not require a pre-determined response label to generate a sentiment score. The two most commonly used libraries are:

1. TextBlob: TextBlob is a part of NLTK library in python. The output of sentiment analysis is a sentiment score ranging from -1 to 1 (which is polarity) indicating how positive or negative they are and a subjectivity score of 0 to 1 where 0 indicates a fact and 1 indicates an opinion. TextBlob finds all the words and phrases that it can assign a polarity and sensitivity to and averages them all together. Finally, each phone is assigned one polarity and one subjectivity scores
2. Vader: Valence Aware Dictionary and Sentiment Reasoner is a part of vaderSentiment library in Python for applying sentiment analysis techniques. The output of Vader method is a compound score metric which is calculated by summing the valence scores of each word in the lexicon, adjusted according to the rules, and normalized to be between -1 (most extreme negative) and 1 (most extreme positive).

Use case: Twitter sentiment analysis for Airlines

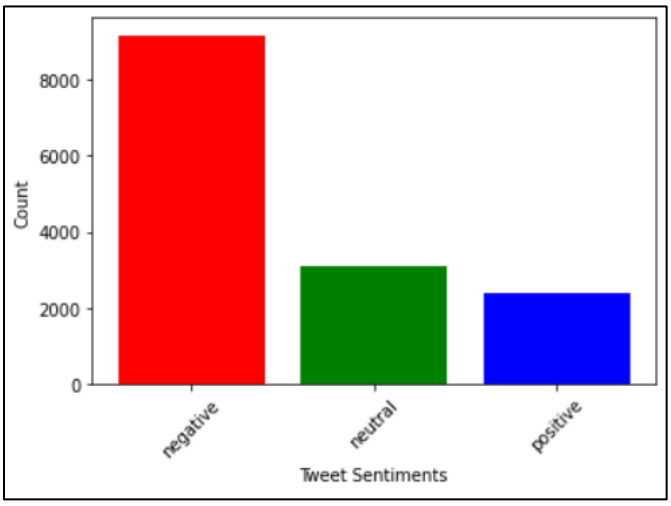
Data Collection

The data for this project was collected from https://data.world/crowdflower/airline-twitter-sentiment and has 14,640 rows and 10 feature variables. It contains Twitter data scraped from February 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

Exploratory Data Analysis

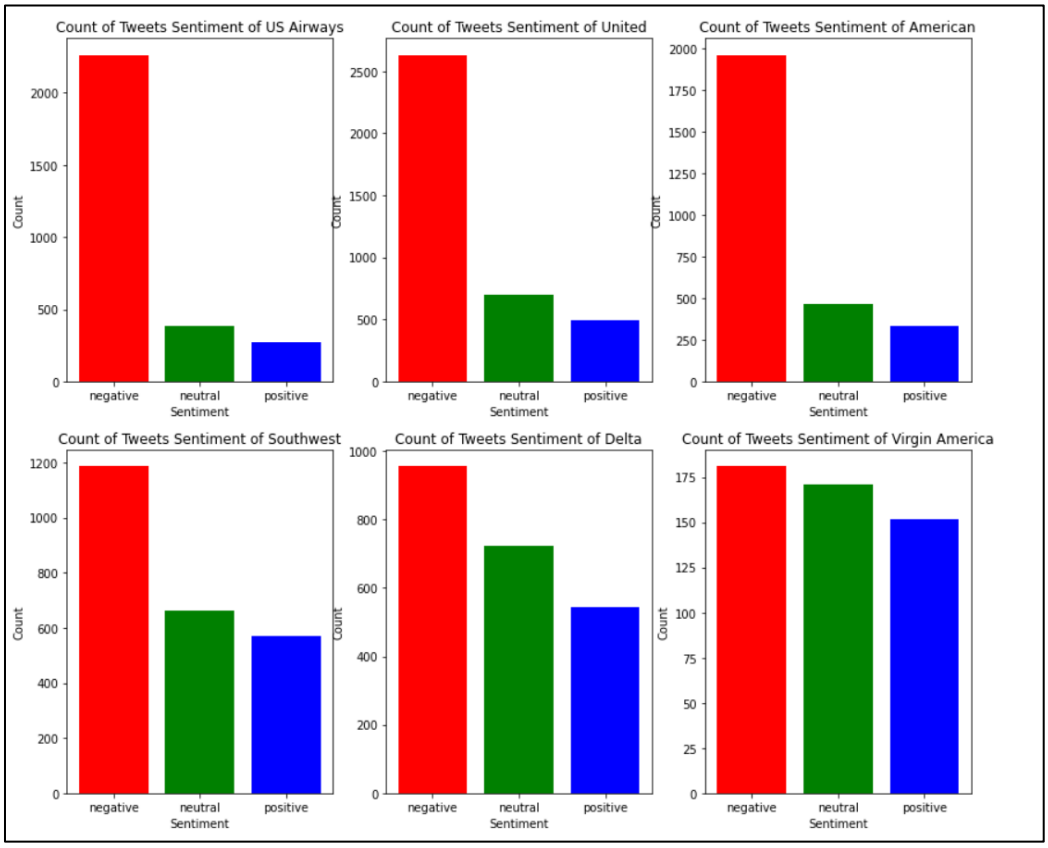
The data contained various fields such as tweet ID, sentiment and confidence about the airline, negative reasons and its confidence, airline, name, tweet text and metadata about the tweet. An initial analysis of the sentiment column showed that there were three categories – positive, neutral and negative, and the negative category had a lot more tweets.

Figure 4 - Tweet counts for each sentiment



The tweets were also segregated by airline category for the sentiment and US Airways, United and American Airlines had most of their tweets in the negative category. Delta and Southwest Airlines were having a few more neutral and positive tweets, but negative tweets were still the highest in number. Virgin America was the only airline that had nearly equal number of tweets in all three categories but was also the only airline with less than two thousand tweets.

Figure 5 - Tweet counts for each sentiment for each airline



Looking at the metadata columns we have the date of creation and geography of the tweet. We look at the positive and negative tweet counts per day per airline next.

Figure 6 - Count of Positive Tweets Per Day

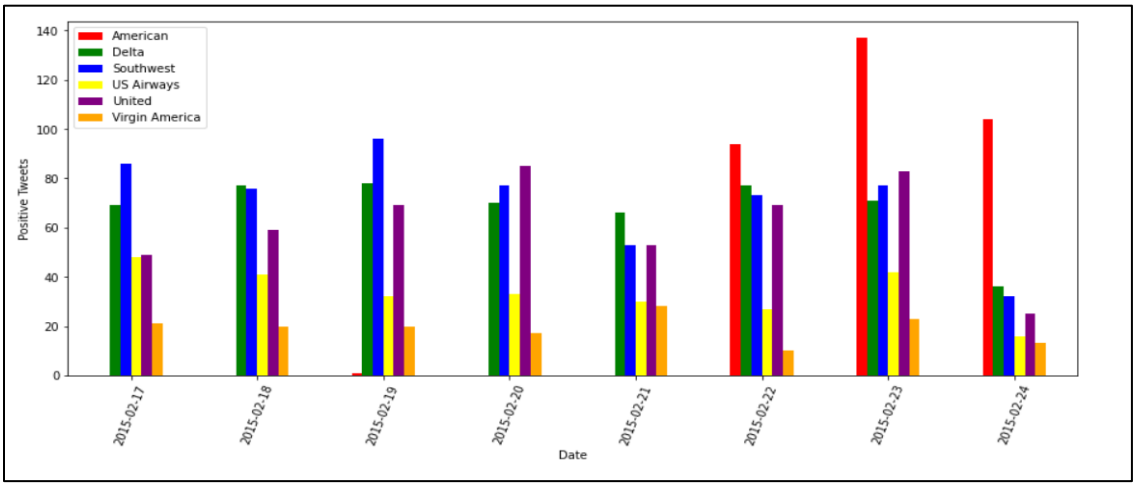
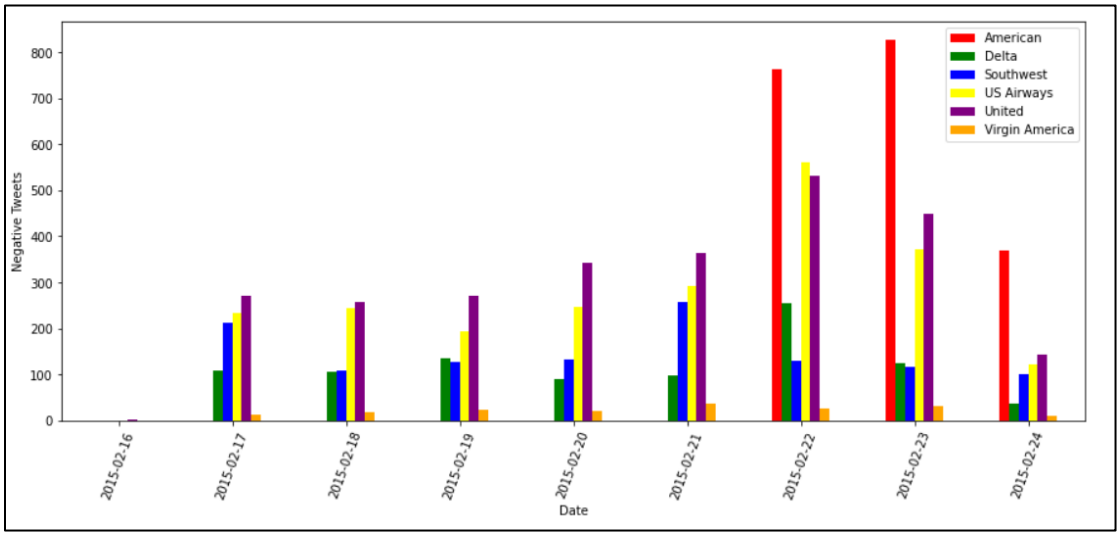
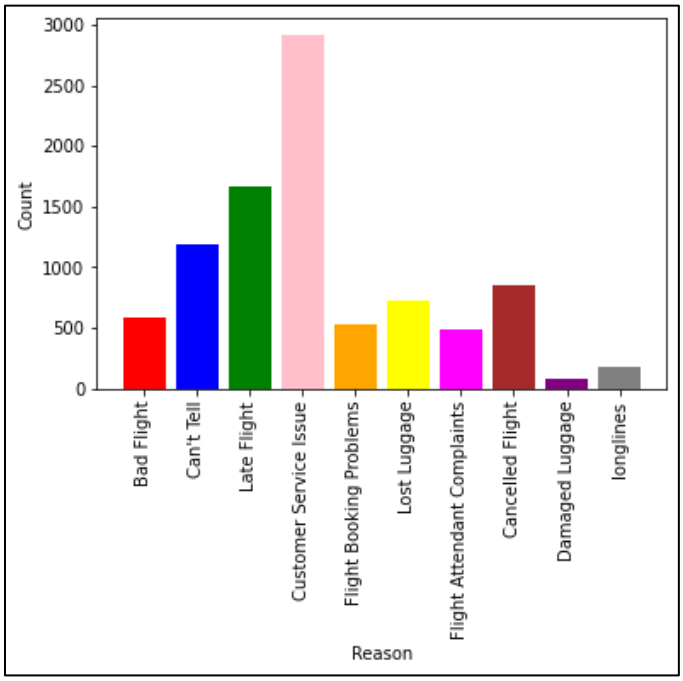


Figure 7 - Count of Negative Tweets Per day



The reasons for negative tweets showed nine categories as shown in the figure below.

Figure 8 - Count of tweets by Negative Reason



Modeling workflow

Text Preprocessing

In order to apply natural language processing techniques, the data was then cleaned and standardized. The following steps were taken:

- Checked for missing and null values in the dataset and removed those columns
- Converted text to lowercase
- Removed punctuations using regular expressions
- Removed stop words and common English words to help us focus on important words
- Tokenized the data to get a bag of words to analyze
- Stemming and lemmatization normalized the data

Using word clouds for both positive and the negative sentiment tagged tweets, we can visualize the word contents. The positive sentiment word cloud had words such as "thank", and "awesome" and surprisingly "customer service" as well. The negative tweets seemed to have the words "bag" and "time" a lot which could be luggage related and flights being delayed.

Figure 9 - Positive Sentiment word cloud

Figure 10 - Negative Sentiment word cloud

Sentiment Analysis

Our next step was to do sentiment analysis based on the sentiment column provided in the data. For this we have two approaches, one using the traditional machine learning algorithms such as logistic regression, K-Neighbours, SVC, etc and the other approach was using packages such as TextBlob and VADER which would take into account the English language rules and assign a sentiment rating.

The response was taken to be binary as either negative or non-negative. The neutral tagged tweets were clubbed with the positive tweets for this analysis. Using this response variable, we then fit the models on a 80-20 split training data and compare the performance on the testing data.

Feature extraction was done by two methods – count vectorizer or bag of words and TF-IDF. The model performances are compared in the table below.

For the packages we looked for the probability of the tweet being negative and compared to a threshold value for assigning it as a negative tweet. The output of this approach was then compared to the original column of sentiment to create a confusion matrix.

Performance Evaluation

We would be evaluating the models based on the following metrics:

- Model Accuracy: The classification accuracy of all classes
- Precision: Percentage of correct predictions
- Recall: Percentage of the class predicted from the total of that class

			Logistic Regression	K-Neighbors	SVC	Decision Tree	Random Forest	Ada Boost	Multinomial Naïve Bayes
Bag of Words	**Model accuracy**		65%	59%	74%	76%	81%	79%	82%
	negative	precision	65%	84%	77%	81%	85%	82%	82%
		recall	100%	45%	40%	81%	87%	86%	93%
	positive	precision	0%	46%	74%	66%	75%	72%	83%
		recall	0%	85%	93%	66%	71%	65%	64%
TF-IDF	**Model accuracy**		65%	73%	66%	72%	81%	78%	79%
	negative	precision	65%	85%	65%	80%	83%	83%	77%
		recall	100%	71%	100%	76%	90%	81%	98%
	positive	precision	0%	59%	100%	60%	78%	68%	92%
		recall	0%	76%	3%	66%	66%	71%	46%

Figure 11 - Model Accuracies with Count Vectorizers

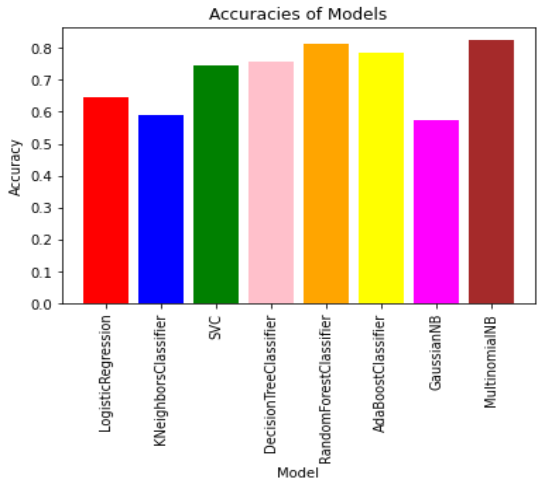
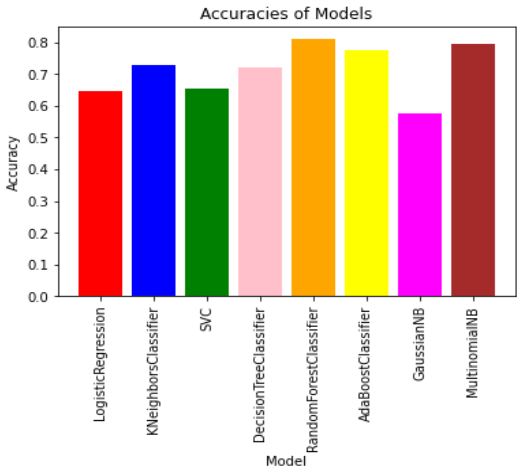
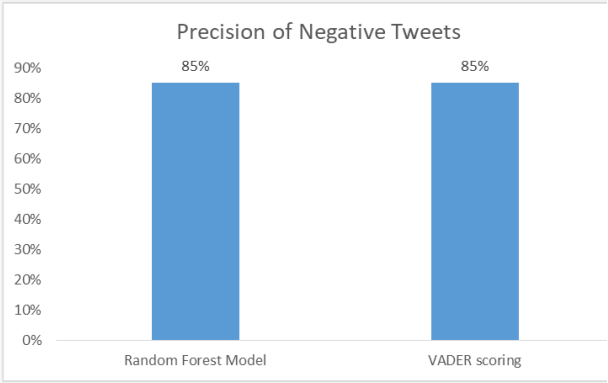
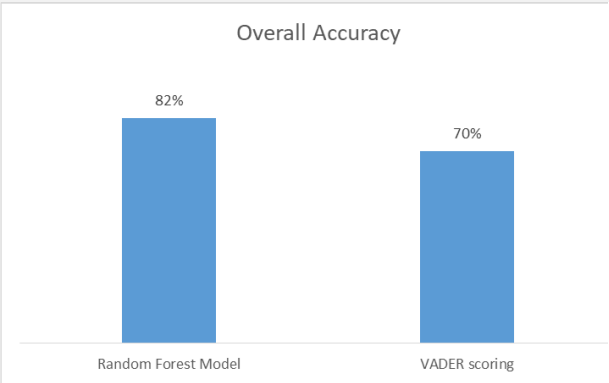


Figure 12 - Model Accuracies with TF-IDF







Findings

Summarizing the insights gained from the machine learning algorithms:

- Bag of words method performs best with Multinomial Naïve Bayes in model accuracy and predicting of negative sentiment tweets:
    - Model accuracy: 82%
    - Negative tweet prediction has a precision of 82% and recall of 93%, both of which are second best among models
    - Positive tweet prediction has a precision of 83% which is the best and recall of 64% which did the worst among other models
- TF-IDF method performed well with good model accuracy using the Random Forest model:
    - Model accuracy: 81%
    - Negative tweet prediction has a precision of 83% which was the second best while the recall was pretty high with 90%
    - Positive tweet prediction has precision 78% and recall 66% which combined was better than other models

Overall, the Random Forest model seems to be a good choice to fit the data and predict sentiments of the tweets. In TF-IDF method it performed the best while it was a close second in the bag of words method.

In the other approach of assigning sentiments using TextBlob or VADER packages, VADER was better at assigning sentiments based on the context of the tweet as it is based on the lexicons of the English language especially social media content.

- It provided a better overall accuracy of 70%
- Negative tweet precision of 85% and recall of 65%
- Positive tweet precision of 58% and recall of 81%

Between the two approaches the traditional machine learning algorithms were still being able to predict the tweets better.

Strategy

After a system for estimating the customer sentiments is established, a dedicated strategy team needs to follow up with a targeted action plan. Macro segments can be created based on sentiment polarity and customer value as shown below. Customer's annual spend (rolling 12 months) can be taken as a measure for customer value.

High Value Negative Sentiment	High Value Positive Sentiment
Low Value Negative Sentiment	Low Value Positive Sentiment

Below we have provided a list of possible strategies for targeting customers prioritized based on the value-sentiment segments

Low value negative sentiments

Priority	Strategy	Components
1	Enhanced Customer Support	Identify key concern areas and triage resolution of outstanding issues
		Continue proactive support for the near future
		Reduced wait times for calling/Jump the queue
2	Customized feedback requests	Targeted feedback requests through email
		Routine pulse checks to amplify voice of customer
3	Customized advertising	Target with ads informing customers about support services and help conduits
4	Small value vouchers/coupons	eCoupons with extended expiry
		Relaxed baggage allowance claims

High value negative sentiments

Priority	Strategy	Components
1	Enhanced Customer Support	Identify key concern areas and triage resolution of outstanding issues
		Personalized email/Call by relationship manager
		Continue proactive support for the near future
		Customized scripts for future tele-calling instances
		Reduced wait times for calling/Jump the queue
2	High value vouchers/coupons	Discounted value added services for the next trip
		Free upgrade coupons (T&C applied)
		Relaxed baggage allowance
3	Enhanced in-flight services	Personalized attention
		Discounted meals
4	Customized advertising	Target with ads informing customers about support services and help conduits

Low value positive sentiments

Priority	Strategy	Components
1	Brand Advocacy	Incentive-linked prompts to provide google/trip advisor reviews
		Incentives for membership referrals
2	Upsell	Promote value added services
3	Customized advertising	Informative ads on new service updates
		Increased exposure to brand awareness promotion ads for sharing on social media
4	Enhanced customer support	Increased email engagement
		Routine pulse checks via telecalling
5	Small value coupons/vouchers	eCoupons with limited claim window
		Relaxed baggage allowance claims

High value positive sentiments

Priority	Strategy	Components
1	Cross-sell/Up sell	Cabin upgrades at reduced rates
		Packaged deals for holidays
		Promote Value added services
2	Brand Advocacy	Incentive-linked prompts to provide google/trip advisor reviews
		Incentives for membership referrals
3	Product offers	Offers on travel advantage credit cards
		Airport Restaurant deals
4	Customized advertising	Informative ads on new service updates
		Increased exposure to brand awareness promotion ads for sharing on social media

Challenges in Sentiment Analysis

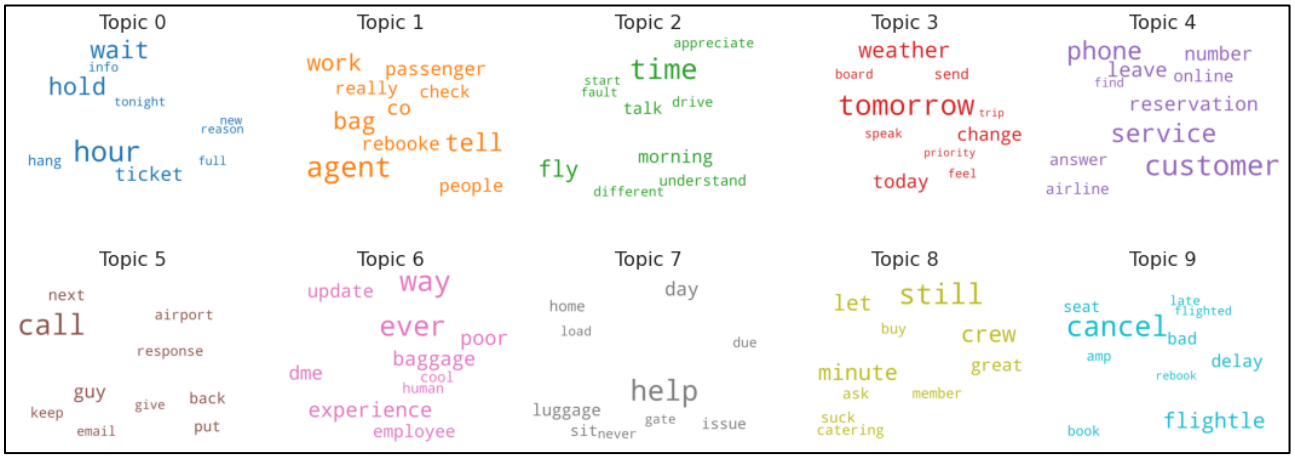
- Irony and Sarcasm - Sarcasm detection in sentiment analysis is very difficult to accomplish without having a good understanding of the context of the situation, the specific topic, and the environment
- Negation Detection - In linguistics, negation is a way of reversing the polarity of words, phrases, and even sentences. Considering negation can significantly increase the accuracy of a model
- Word Ambiguity - The problem of word ambiguity is the impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context
- Multipolarity - Sometimes, a given sentence or document—or whatever unit of text we would like to analyze—will exhibit multipolarity. In these cases, having only the total result of the analysis can be misleading, very much like how an average about outliers
- Repeated letters People often repeat letters in some words, in order to stress upon an emotion. For example: sad, saaaad, saaaddd. All of them mean the same, yet it is not possible to distinguish between them if guided only by their spellings
- Hashtags Words in hashtags may be read different from the same word without the hash tag

Further scope

Having identified the customers with positive or negative sentiments, the next step is to find and address the topics of concern or appreciation. We can leverage topic modeling techniques to extract the hidden topics from the text, enabling airlines to find naturally occurring topics on unsegregated data. Airlines can then use this in categorizing customers into different market segments for ad campaigns, offers and other customer acquiring strategies.

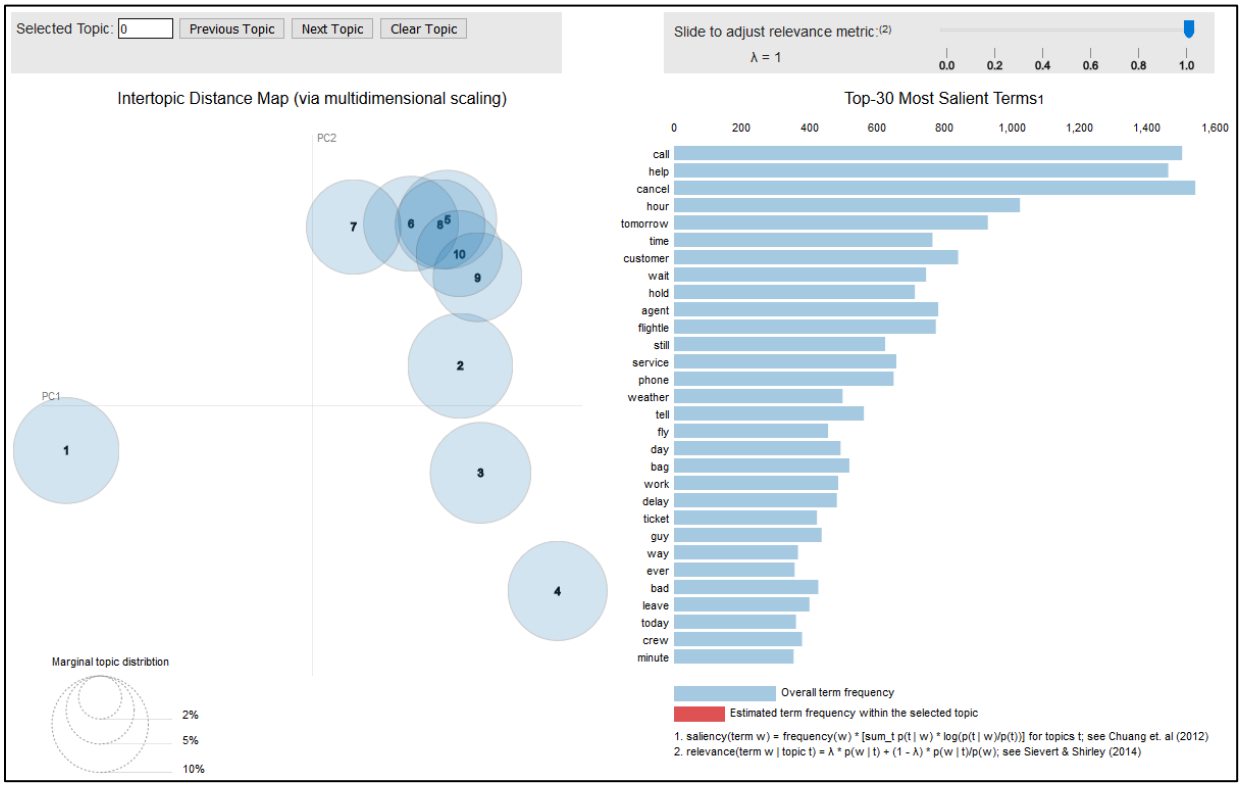
For illustration purposes we picked the negative tweets and attempted to classify the text corpus into 10 clusters using the Latent Dirichlet Allocation (LDA) in Python's Gensim package. The word cloud below shows the mix of words for topics. For example, the Topic 0 in the word cloud below was inferred to be related to information services where the words "wait" and "hold" indicate long calls waiting. Similarly, Topic 3 was found to discuss the weather before flight departures and Topic 7 discussed the luggage which had the keyword "help" indicating that there are tweets that talk heavily about it.

Figure 13 - Topic Modeling word cloud



A better visualization is using the pyLDAvis package in python which provides saliency and relevancy of each keyword in the corpus. The figure below illustrates this with the topics as circles whose sizes indicate their relative importance and distance indicates the level of separation of the topics. This interactive visualization allows the user to click each topic circle and analyze the top 30 keywords of the topic.

Figure 14 - pyLDAvis for topic modeling



Clustering text data into good quality topics which are clear, segregated and meaningful require context-based text preprocessing steps and finding the optimal number of clusters.

Acknowledgements

We would like to acknowledge Professor Roger Chiang of the Department of Operations, Business Analytics, and Information Systems at the Carl H. Lindner College of Business, University of Cincinnati for providing us with the opportunity to learn and implement a machine learning exercise using natural language processing.

We also acknowledge the repository data.world for enabling data enthusiast all over the world to access open source datasets.