MGMT 4084 – Digital Media Analysis

Professor: Richard Boire



Airline Project – Analytics Phase

Leveraging Twitter Data to Understand Public Sentiment and Opinion about the Airline Industry amid Coronavirus Pandemic

Prepared By: Mirza Razin Alam

**DATA COLLECTION**

While attempting to understand public sentiments and opinions on the airline industry during the covid-19 pandemic, we selected a couple of key sources where we would extract and examine relevant data. For our primary data we used python to extract tweets from the website twitter, using publicly available tweets to help outline some of the problems and patterns people were expressing about the airline industry during the pandemic. Our secondary data source was Mckinsey & Company, a consulting firm which specializes in market research and analysis of big data. Each of these sources proved useful in their own way; with our twitter data we could create some assumptions and key findings and with our secondary data we could compare our results to see if there were any correlations in the data.

From Twitter we looked for tweets that matched our specified keywords. More specifically, while collecting data we used the keywords ‘Flight’ and ‘Covid19’ to look for significant tweets which related to our topic of interest. By using these inputs, we were able to scrape data that was directly related to the Covid-19 Pandemic and the airline industry.

We were able to separate the tweets utilizing a tweet scraping tool called TWINT which is written in Python and store them as plain text. The scraping tool collected data from tweets by returning a dataset with 20,034 entries, which initially included 39 variables by default. This was only the first step in creating the raw dataset as a csv file. Of the 39 variables only 10 were used for this analysis: id, created\_at, date, tweet, language, hashtags, cashtags, user\_id, urls, and retweet.

**DATA PROCESSING, TRANSFORMATION, AND EXPLORATION**

The data processing, transformation and analysis were all carried out in Python using its diverse set of modules, these include pandas, numpy, sci-kit learn, matplotlib, seaborn and more. We loaded our csv file extracted from TWINT onto python to carry out data cleaning techniques, we began by removing duplicate entries and dropping columns that had no value for the analysis. Next, we removed groupings of insignificant words which appeared throughout our data, as well as filtering out non-English words from the texts.

To remove the unnecessary words from the data we used a Natural Language Processing tool called NLTK library in Python, this function filters out all the ‘STOPWORDS’, which are the common English words which do not add value to the data. The stop words needed to be removed because they will impact the quality of the text frequency analysis. Some of the stop words are prepositions and pronouns such as ‘i’, ‘we’, ‘and’, ‘is’, ‘myself’ etc. We created a new column within the dataset called ‘cleaned\_text’ and with the tweets without the ‘STOPWORDS and other unnecessary texts. We concluded our data cleaning by filtering out the tweets that were not in English, and our final dataset was left with 14,287 entries. Once the data is processed into a clean dataset, we were able to conduct the analysis in a much more efficient way.

**TYPE OF ANALYSIS**

We began our analysis by using a data collection technique called web scraping which would allow us to extract data from Twitter using an advanced Twitter scraping tool, written in Python, called Twint. Twint was used to scrape Tweets from Twitter profiles without using Twitter’s API. The main purpose of using this method is to conduct sentiment analysis, which is the process of analyzing whether a piece of writing, such as customer and public comments and reviews, are positive, negative, or neutral.

Sentiment Analysis is conducted by using Natural Language Processing (NLP) methods and machine learning algorithms, and it allows those analyzing data to measure public opinion and classify the sentiments that are expressed into polarities – positive, negative, or neutral. Being that there are different types of Sentiment Analysis, it is necessary to clarify that we are hoping to conduct fine-grained sentiment analysis, which will allow us to group sentiments into more nuanced polarities (i.e. “very positive”, “positive”, “neutral, “negative, and “very negative”).

Based on the frequent words from the texts within the dataset, we assigned 5 topics that are relevant to the airline industry and Covid19. Machine learning algorithms allowed us to classify each tweet from our dataset and by label them by one of the assigned topics. We were also able to carry out a sentiment analysis on these 5 topics. A correlation matrix generated through Python modules also gave us further insight into the relationship between the topics.

**KEY FINDINGS**

**Word Frequency Analysis:**

We were able to investigate the most used words in the tweets that are related to the airline industry and Covid19. This should give us an idea of how the public is reacting to taking air flights during Covid19. Using Natural Language Processing we can analyze each distinct text to create a word cloud (Appendix A). A word cloud is a great visualization for representing the most frequent words in a text file. In the word cloud the most occurring words will be displayed with a larger font.

From our extracted data we were able to identify the most frequent words in our word cloud, such as ‘airline’, ‘case’, ‘people’, ‘vaccine’, ‘pandemic’, ‘covid’. The words with the highest frequency may not particularly be related to the airline industry, however we could also find words such as ‘airport’, ‘flying’, ‘travel’, ‘cancelled’, ‘restriction’, ‘booked’, ‘passenger’, ‘ticket’, ‘trip’ that are more associated with the airline industry. From words like ‘flying’, ‘travel’, ‘cancelled’, ‘restriction’ we can assume that these are among the most frequent discussions related to the airline industry during the pandemic.

**Sentiment Analysis Using Subjectivity and Polarity Scores:**

Using the insights from our sentiment analysis, we can come to conclusions on what is the current sentiment or feeling of the public is towards the airline industry. Sentiment analysis using modern tools that can significantly ease the burden of airline companies by analyzing thousands of references across different mediums.

Textblob is a python module that lets us conduct sentiment analysis by analyzing the text data and classify them based on subjectivity and polarity. Again, to reiterate, the subjectivity score lies between 0 and 1, where a score closer to 1 means the tweet is an opinion rather than a fact. Similarly, the polarity score lies between -1 and 1, where a score closer to 1 is considered positive, a score closer to 0 is considered neutral and a score closer to -1 is considered negative. For our dataset we were able to construct a scatter plot for visual representation showing the subjectivity metric on the Y axis and the polarity metric on the X axis.

We can see that many of the tweets are within the polarity range of -0.25 and 0.50 (Appendix B). This tells us that most of the tweets carry either neutral or positive sentiment. Extreme cases with scores -1 and 1 are quite low. We were able to count the number of tweets and classify them whether they are positive, neutral and negative. The results show that 6047 tweets were positive, 5737 tweets were neutral, and 2503 tweets were negative (Appendix D). Our initial assumption was that there be a higher percentage of negative tweets related to airline industry during the pandemic, however that was not the case.

For subjectivity, the scores vary heavily between 0.2 and 0.8 in the scatter plot (Appendix B). Extreme cases with subjectivity scores of exact 1 or 0 is very low. The results show that 9375 of the tweets are classified as opinions and 4912 tweets are classified as facts (Appendix C). So, most of the tweets within our dataset are identified as opinions by the NLP algorithm.

**Topic Analysis:**

Topic analysis helped us in studying certain topics from a wide number of tweets using a machine learning algorithm that identified each tweet based on our specified topics. Such a topic wise analysis is frequently used as a text mining tool to reveal semantic structures within a wide number of texts. We picked **Airport, Travel, Passenger, Safety and Restriction** for our analysis as these are some of the frequent words which are relevant to the airline industry and that has helped us in understanding the attitudes of the people towards the industry.

We found that out of more than 14,287 tweets that we could gather, the maximum number of tweets were related to passengers, which comprised of upwards of 4500 tweets, this was followed by Travel, which coupled with around 3500 tweets (Appendix F). Almost 2500 tweets were related to Airport and rest of the tweets were classified by the topics safety and restriction which was less than 2000 for each.

From our sentimental analysis of these topics, we saw that airport, travel and passenger shared similar sentiments. We found that the sentiment behind the tweets on these topics was mostly either neutral or positive (Appendix E).

The analysis of the topic of safety shows the least amount of negative sentiment as people understood its importance and prioritized safety during these times and therefore safety had a majority neutral or somewhat positive sentiment relating to it. People were mostly in favor of the restriction as the need for the hour and they seem to have a positive outlook towards the topic, however among the other topics, restrictions also had the highest percentage of negative sentiment in our study.

We grouped together the terms showing similarity, based on the correlation between those words in the tweets. We conducted correlation analysis (Appendix G) on the topics **Airport, Travel, Passenger, Safety and Restriction** to find patterns in the relationships between the topics. We saw that Travel and Restriction were positively correlated with a coefficient of 0.65 which is the highest among the topics, Airport and Travel has correlation coefficient of 0.51 whereas Safety and Travel had 0.56 between one another. Similarly, Safety and Airport were correlated by 0.46 and restriction and safety have coefficient of 0.52. The topic Passenger has a weak positive correlation coefficient with all the other topics. None of the topics had negative relationships with one another.

**Conclusion and Recommendations:**

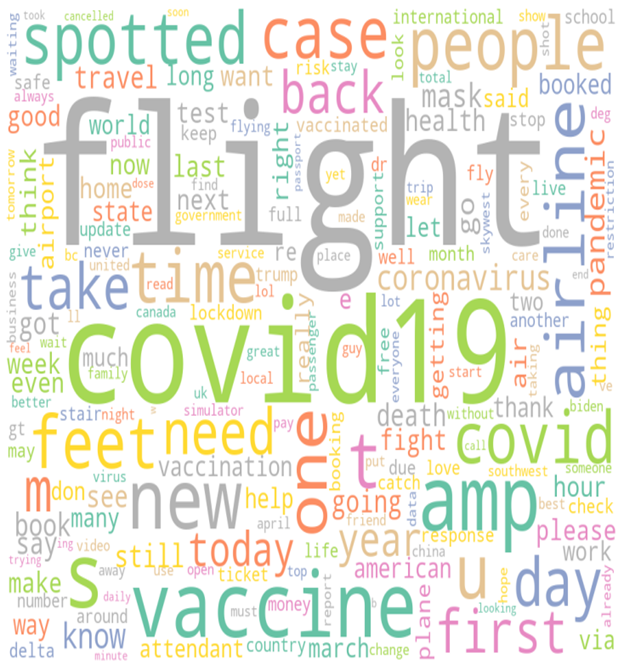
From our analysis based on the 14,287 tweets, we can conclude that most of the tweets are either positive or neutral in terms of public sentiment. This was surprising to see as we had expected a larger percentage of negative tweets. This is favorable for the airline industry as it had suffered heavily during the start of the Covid19 pandemic.

From these opinions of the public, it will be easier for airlines to understand about negative and positive opinion of people and can be figured out to enhance airline services in the future considering how rapidly the industry was affected by the Covid19 pandemic.

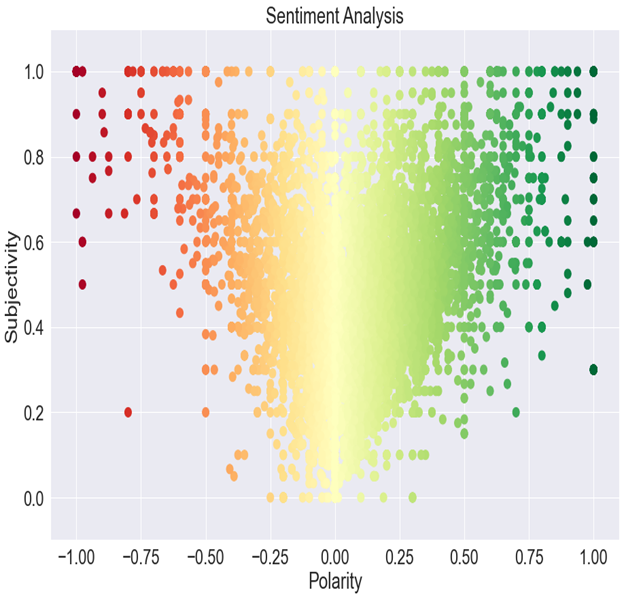
The analysis from the current data is however, not statistically significant as the sample size is very small compared to the population. Further research is needed with a vast to find out how sentiment has shifted since the beginning of the pandemic which could potentially include millions of tweets.

**APPENDICES**

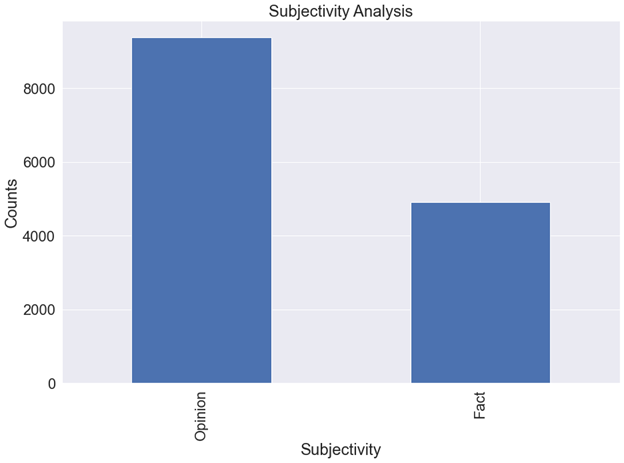
**Appendix A: Word Cloud**



**Appendix B: Sentiment Analysis with Subjectivity and Polarity Scores**



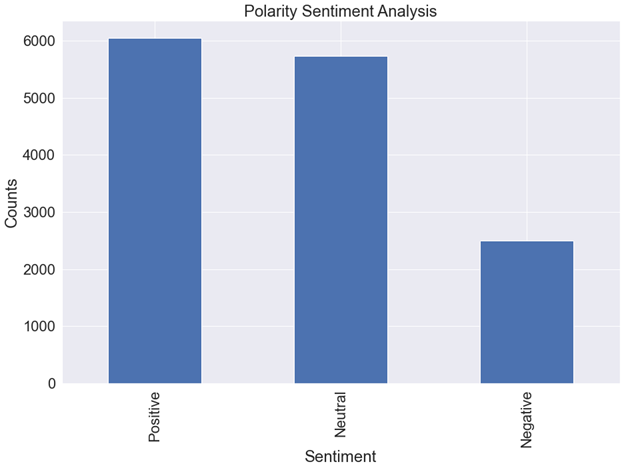
**Appendix C: Number of tweets based on Subjectivity**



Opinion 9375

Fact 4912

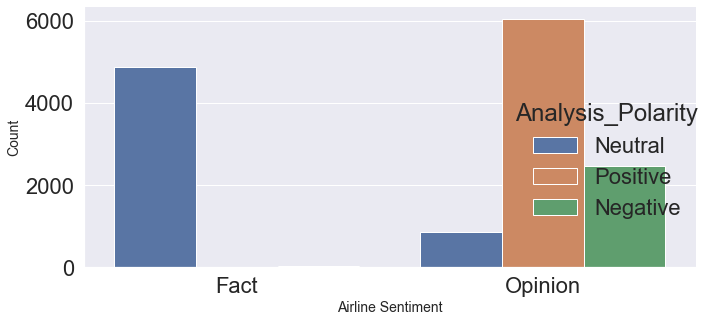
**Appendix D: Number of tweets based on Polarity**



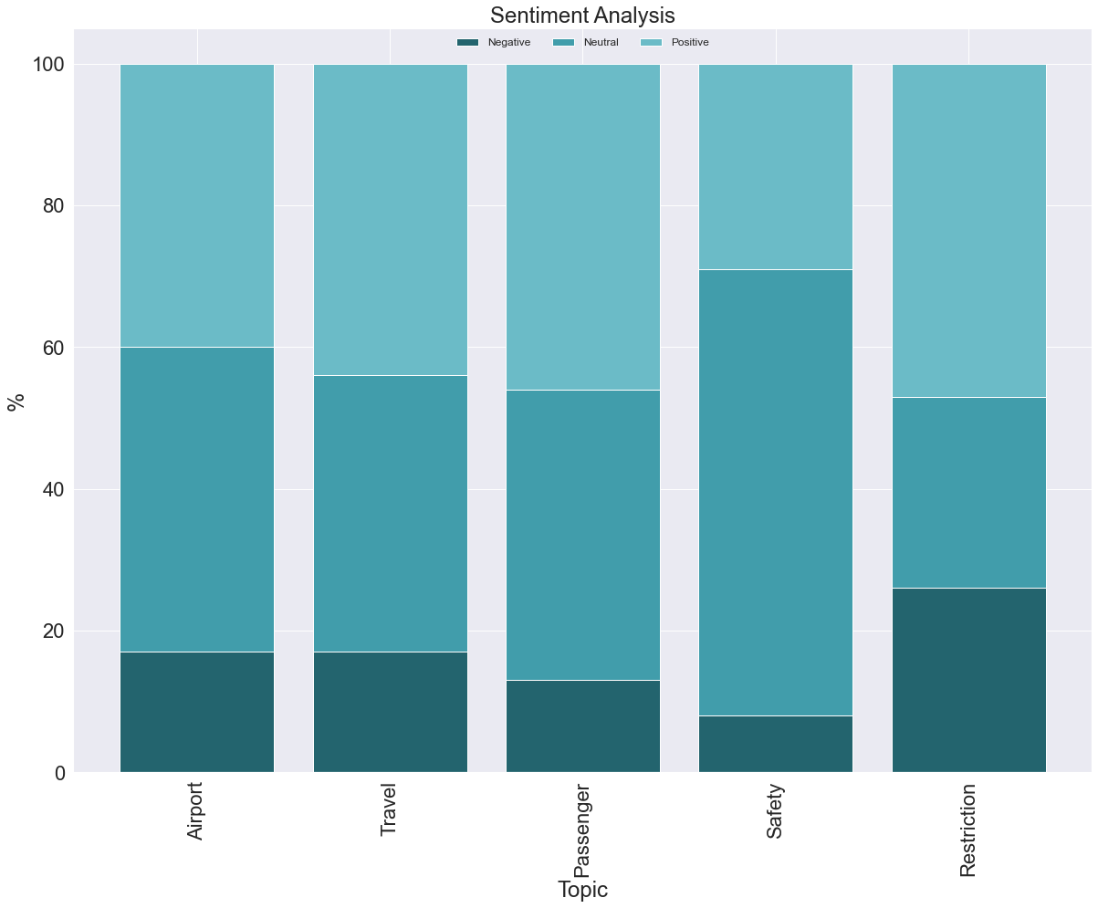
Positive 6047

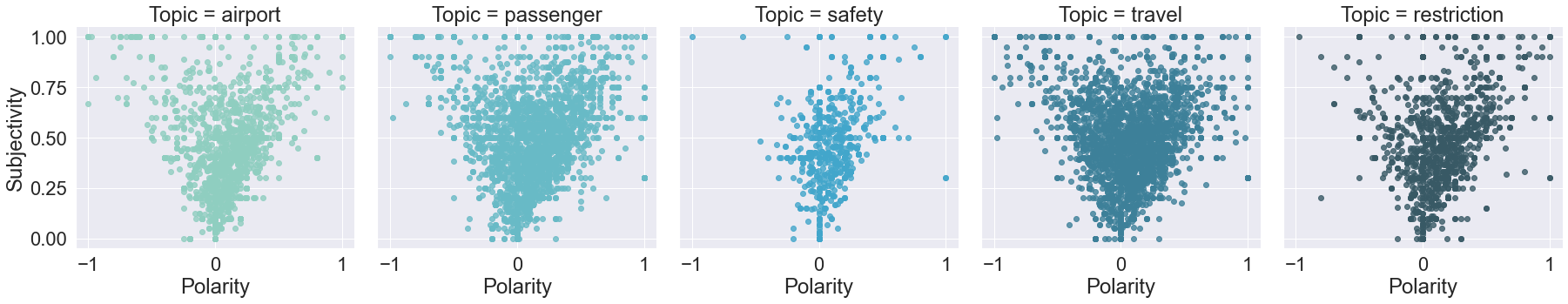
Neutral 5737

Negative 2503

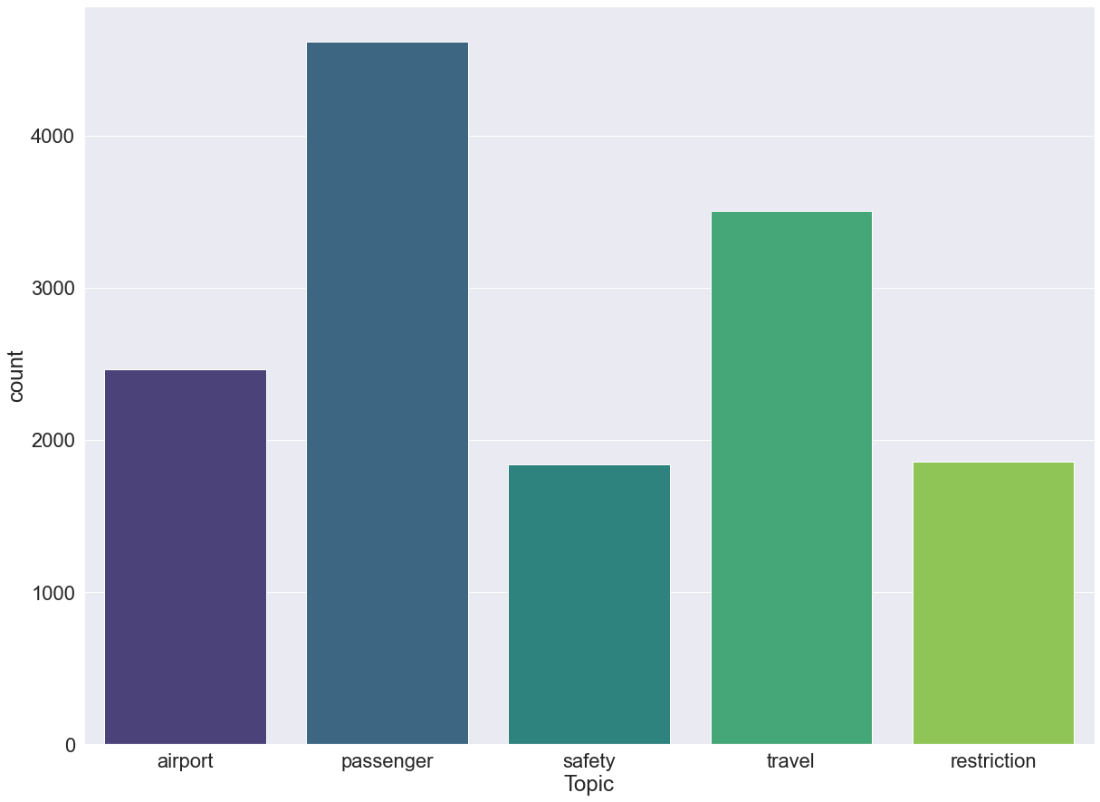


**Appendix E: Sentiment Analysis on Topics**





**Appendix F: Number of Tweets Classified by Each Topic**



**Appendix G: Correlation between Topics**

