Transit Related Tweets Analysis in New York City

Chang Du, Wenjie Zheng

DS-GA 1015 Text as Data

Word Count: 3116

Abstract

The rise of social media platforms including Facebook, Twitter, and Instagram provides a whole new platform for social science researches in a fashionable manner. The transit agencies recognize the convenience and information accessibility to the greater audience population by the social media platforms and thus start using social media as information dissemination and advocacy tool. In recognition of the crucial functions of social media, it is intriguing to analyze the interaction between the general public and transit service providers on the social media platform, and, in this case, Twitter. This paper conducted a sentiment analysis with both text and emojis from the Tweets for ‘MTA’ related tweets and identified the hotspot stations that are prone to infrastructure malfunctions, delays, and other factors that lead to service interruptions. Overall, the sentiment analysis suggests that the positive attitudes outnumbered other attitudes and successfully identify the service hotspots from the textual information.

Keywords: Twitter, Hashtag, Tweets, Transportation, Social Media

# Introduction

Performance measurement has been one of many methods evaluating the service delivered by the transit agencies. These measurements often include timeliness of the schedule, safety and security, comfort and convenience, cleanliness and sanitation of stations and onboard, and courtesy, etc. (MTA, 2012) The Metropolitan Transportation Authority, commonly known as MTA, runs the largest mass transit in the United States including the conglomeration of buses, subways, Metro-North, and Long Island Railroad, and among the top of the world in terms of annual ridership, system track mileage, and the number of stations, not to mention that the system runs 24-7. It is one of the oldest public transit systems as well. However, years of maintenance negligence, aging infrastructures, and threats of natural disasters hinder the day-to-day operations of the system. The MTA Subway Performance Dashboard has shown that there is a steady decreasing trend on the statistics of the major incidents which result in a major delay that influence more than 50 trains. (MTA, 2018) Other measures including terminal on-time performance, subway car breakdown rate, and extra platform wait time are also seeing positive trends since the overall performance hit an all-time low in January 2018. The on-time performance has been steadily increasing ever since, which improved from 58.1% to 78.3%. (Ibid) We can learn that the transit and commute have been significantly improved out of the MTA statistics posted. Despite the positive trending on these performance metrics that come from single-handed MTA statistics, it is also important and intriguing to learn what the riders think and feel about the day-to-day commute on the MTA. One of many approaches is to go through the numerous Tweets people post every day and figure out what peoples’ opinion about their MTA rides are. The percentage of the population in the U.S. that use any social media platform has risen to roughly 80% from 10% in 2008. (Edison Research; Triton Digital; Salesforce.com, 2019) Some people even spend more time on social media than other activities and getting information and news from the social media platform is one critical source other than traditional sources such as newspapers, magazine, and television. Transit agencies have also heavily used social media means including Twitter to inform the public about service updates, crew appreciation, and board meetings, etc. Riders also used Twitter to express their opinions on transit experience. In this way, we are expecting some candid and unfurnished points of view. What’s more, in recognition of the nature of the Twitter posts about MTA, the analysis is expected to uncover both praise and blame from both the context and emojis. Yet in keeping an unbiased mind the paper seeks to find out what aspects people are satisfactory about and what aspects MTA needs to work on and improve, which helps MTA better allocate resources and better address issues in a timely manner.

# Literature

Relevant researches and papers provide crucial contextual knowledge in helping us better approach our objectives and understand the use of social media in transit-related news. To begin with, Cottrill examined the used of Twitter in large events. (Cottrill, et al., 2017) Social media information about transportation comes in a timely and cost-effective manner for transit agencies to examine and study. This can be especially helpful during the peak hours and emergencies which resources and strategies can be deployed quickly to address urgent issues. The transit system is often built and operate at the capacity of rush hours and off-peak hours when there is a particular large number coming in and out of the transit system during a short period of time in the morning and evening and steady off-peak passenger volume throughout the day. In the case of a large event like sports game and concert which the relevant transit options will face the extremely high transit demand and potential large passenger influx into the subway stations and bus stops, social media is a crucial platform for transit-related information and guidance. This article defined the aforementioned phenomenon as ‘transport disruption,’ and did a case study on the Twitter *account @GamesTravel2014* to evaluate how this account is used throughout the course of 2014 Commonwealth Games in Glasgow, Scotland. The paper explored the use of the official Commonwealth Games account for the dissemination of transport information that involves multiple transit agencies in the city of Glasgow, which include First Bus Glasgow, ScotRail, and Glasgow Subway, etc. The Tweet analysis examined the number of activities including direct comment, original tweet, and retweet on a number of transit service providers. The official game account retweet transit agencies’ Tweets for better dissemination of transit information. Dimitrios and Constantinos examined the use of social media platform for transportation data collection and survey by constructing a web-based questionnaire survey that asks people about carsharing and bike sharing using google forms. (Dimitrios & Constantinos, 2012) This survey is not limited to Greek recipients, but to an international community. The research team compared the feedback from different means of surveying, including social media and internet, and mailing. The survey conducted through internet dissemination receives a much larger response compare to mailing, yet the demographics results turned out to be somewhat biased since the 90% of the responders are between the age of 18 and 35. The research then recognizes the importance of Twitter as a crucial platform for people to socialize due to its popularity and a large number of Tweets posted daily, not to mention the Twitter as an essential data storage for researchers to look into the ‘attitudes and perceptions of the general public, as well as information on emerging incidents.’ The Collins’ group did rider sentiment analysis with Twitter data for Chicago Transit Authority CTA. (Collins, Hasan, & Satish V. Ukkusuri, 2013) The analysis used an algorithm called SentiStrength for sentiment strength detection. It contains an original collection of 298 positive terms and 465 negative terms and has a scale of 1 to 5 for positive sentiment and -1 to -5 for negative attitudes. The research team manually collected tweets containing the keyword relevant to the Loop ‘L’ subway lines which operate under the division of CTA. 557 pieces of Tweet texts were collected to make up a large enough dataset for the SentiStrength model. The research also utilized Twitter’s Streaming API to monitor the volume of tweets in real time throughout the day. The result was presented in plots that show the variation in sentiment scores of transit-related tweets in aggregated statistics. Overall, the research team recognizes the strength of analyzing transit-related Tweets that data can be collected in real time and user-specific needs can be assessed.

# Theory and Hypotheses

This paper provides a preliminary summary of the Tweet texts and does sentiment analysis on transit-related Tweets. What’s more, the paper also provides a methodology for identifying hotspot stations and lines that are prone to service interruptions.

# Data and Methods

## Data Sources

As the analysis is primarily conducted on the subject of Twitter data, the scraped tweets are the primary data source. Several separate Twitter accounts were taken into consideration for the scope and their tweets and posts were being acquired in the data collection process. There are lots of transit-related Twitter account available for analysis in Greater New York Area including the Metropolitan Transportation Authority (MTA), New Jersey Transit (NJT), Port Authority of New York and New Jersey (PANYNJ), and several subdivision agencies of MTA including New York City Transit Subway, New York City Bus, Long Island Railroad, Metro North, etc. Each of the transit service providers is in charge of different area and field as their name suggests. To narrow down the scope, this study focused on Twitter accounts include ‘@MTA’ and ‘@ NYCTSubway’ since they are some of the most well-known agencies. The transit-related hashtags were considered to be able to provide riders’ expressions and feelings on the MTA and subway. These hashtags include ‘#mta’ and ‘#nycsubway’ which people often tweet about. The data from the official account timeline ‘@NYCTSubway’ was also collected in order to analyze the stations and lines that are prone to service interruptions for the past three weeks. To acquire tweets, a Twitter account was registered and the Twitter API Key set was obtained after filing for a Twitter API application. (Desai, 2018) 10,000 Tweets were expected to be returned after each API search. However, the number of transit-related tweets did not exceed 10,000 pieces in any single search. Instead, the dataset is rather much smaller and all the tweets in the past 7 days at the time of the search were collected. The MTA official account timeline only tweeted 90 times in the past few years, while the ‘@NYCTSubway’ account timeline has roughly 1,500 records in three weeks, and the MTA hashtags have roughly 3,700 records for 7 days. The account’s timeline was collected by using the *userTimeline()* command in the *twitteR* package and the keyword hashtag was collected by using the *searchTwitter()* command. In order to increase the size of the dataset, the Twitter API search on the aforementioned Twitter account and hashtags were conducted three times on April 21st, April 28th, and May 10th. What’s more, the emojis in the tweet text was also a great subject for sentiment analysis since they may show the feelings directly. Since the emoticons are displayed inproperly in R, we need to decode them in the MTA tweets data before analysis. The emoji decoder obtained contains information for 842 emojis*.* (Peterka-Bonetta, 2015)The sentimental context data of emojis was obtained from Emoji Sentiment Ranking 1.0**,** which includes 969 emojis. (Kralj Novak, Smailović, Sluban, & Mozetič, 2015) Besides the Tweets dataset, a CSV file of subway station information in New York City was collected from the New York City Open Data Portal that provides the names and geography coordinate of stations and the subway lines that serve the particular station.

## Data Examination and Integration

Given the datasets and sources mentioned in the table above, every dataset is then downloaded individually and cleaned individually before being analyzed.

For each of the aforementioned datasets, three files were merged into one with *rbind()* and ordered by the time they are posted in terms of date-time. 8 Columns were selected out of the 17 columns, including ‘text,’ ‘favorite count,’ ‘created,’(date-time) ‘retweet count,’ ‘longitude,’ and ‘latitude’. There were some duplicates in tweet records due to multiple times of data collection. For example, the dataset collected on April 28th contains some tweets posted on April 20th, which were also contained in the dataset from April 21st. Thus, the duplicated entries were deleted when two entries have the exact text content and the date-time they were tweeted. Upon completing the initial data examination, cleaning, and integration, the riders’ opinion-related ‘#mta’ hashtag data frame has 3,784 rows, the ‘#nycsubway’ hashtag has 832 rows, while the subway service updates related ‘@NYCTSubway’ timeline data frame has 1,490 rows.

As for the subway station information dataset from the NYC Open Data Portal, the subway line names and subway station names are needed. There are 473 stations in the dataset while the stations along the Staten Island Railroad (SIR) are not included in this dataset. Each record has the station ID, station name, GIS coordinate, line service, and notes that indicate service coverages.

**Preprocessing**

The preprocessing of the tweet text, in fact, has been the most crucial and most time-consuming part of this study. The preprocessing steps are done in a slightly different manner for the hashtags and the subway timeline since the hashtags are mainly used for the sentiment analysis while the subway timeline is used for hotspot analysis.

To begin with, the ‘#mta’ and ‘#nycsubway’ texts were all lower-cased and the punctuations and English stop words were removed. Besides removing various signs, special characters, and website URLs, the ‘rt’ characters were also being removed. Some tweets starting with ‘rt:’ suggest that they are the replies of other tweets, but ‘rt’ words themselves do not add any meanings or contexts to the tweet analysis.

In comparison to basic preprocessing done to the hashtag dataset, the preprocessing for the ‘@nyctsubway’ and the subway station information dataset got much more complex. The analysis processed later showed that the station names have to exactly match each other to be identified. The ‘th,’ ‘rd,’ and ‘nd’ characters were substituted with blank spaces with *gsub()* function while all the ‘ave’s were substituted with ‘av’ after recognizing that the ‘@nyctsubway’ tweets ‘av’ instead of ‘ave’ for station names ending with ‘ave.’

**Analysis**

In the first step, the *textstat\_collocations()* command identifies the significant collocations (bigrams) in the text that appear relatively many times. The ‘#btrain #dtrain’ appear the most with 238 counts followed by ‘#4train #5train’ with 210 counts and ‘#rtrain #wtrain’ with 191 counts, suggesting that these line pairs appear often and tend to appear together with each other in the tweets. (Figure 1) The word cloud of ‘#mta’ related terms shows that ‘delays,’ ‘service,’ ‘work,’ ‘change,’ and etc. have high term frequency. (Figure 2) Tweets with ‘#mta’ get 1.06 favorites and 766 times of retweets on average for the period between April 14th and May 10th. The textstat\_collocations() and word cloud on ‘#nycsubway’ showed similar results, several subway lines have high frequencies including ‘#btrin #dtrain,’ and ‘#4train #5train.’ (Figure 3) (Figure 4) Words like ‘waiting,’ ‘investigation,’ and ‘homeless’ also make the list by having at least 20 appearances.

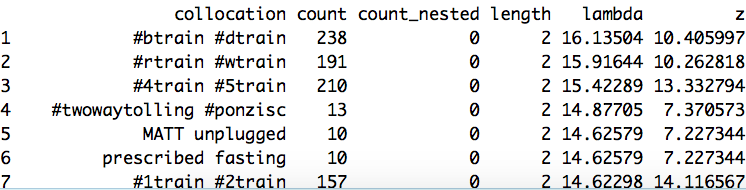


Figure #mta Bigrams

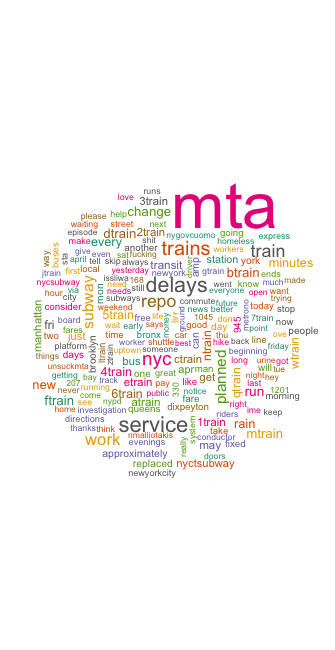


Figure #mta Word Cloud

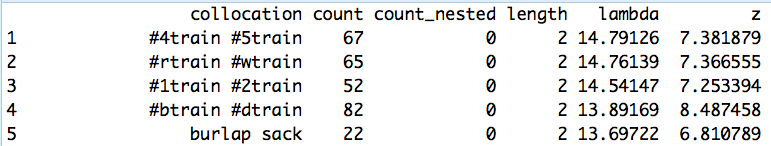


Figure #nycsubway Bigrams

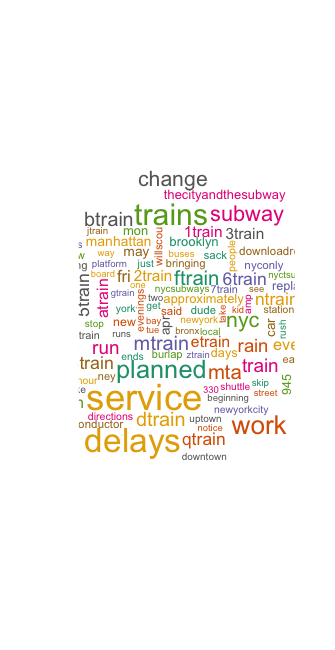


Figure #nycsubway Word Cloud

The Term Frequency-Inverse Document Frequency with *topfeatures()* showed that ‘delays’ and ‘fixed’ are among the top terms in the text along with ‘mta,’ ‘service,’ ‘nyc,’ and ‘train.’ Sentiment analysis was also done to the text with Hu & Bing Sentiment Words Dictionary. If there are more positive words in a tweet text, a positive sentiment score will be assigned. Wice versa, a negative word will be assigned if the text has more negative words. If there are equal number of positive words and negative words, the text will receive a sentiment score of ‘neutral (0).’ The statistics have shown that the tweets with negative moods are 26% more than tweets with positive moods while 40% of all tweets do not show strong positive or negative opinion. (Figure 5) Similarly, there are 47% more negative-mood tweets than positive-mood tweets for ‘#nycsubway’-related tweets. (Figure 6)

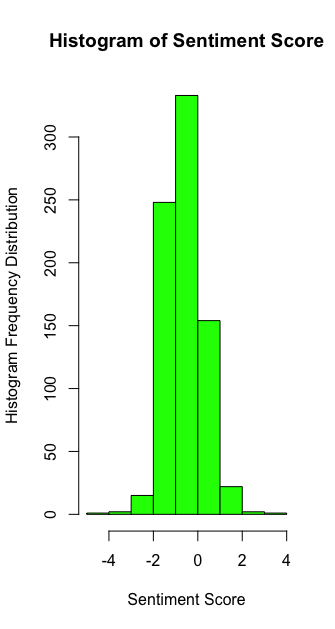
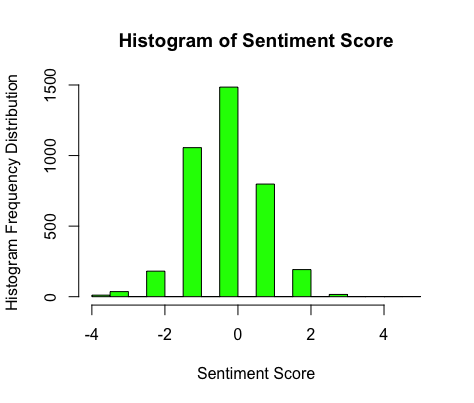


Figure #mta Tweet Sentiment Score Histogram Figure #nycsubway Tweet Sentiment Score Histogram

The NRC Sentiment Analysis is able to process the text and compute of a statistics of the frequency of different words that are associated with ‘anger,’ ‘positive,’ ‘sadness,’ ‘joy,’ ‘trust,’ ‘disgust,’ etc. (Desai, 2018) For the NRC score of ‘#mta’ tweets, there are more words about ‘positive’ and ‘negative’ followed by ‘trust’ and ‘anticipation.’ (Figure 7) As for the NRC score of ‘#nycsubway,’ there are more words about ‘positive’ and ‘fear’ followed by equal amount of ‘negative’ and ‘trust.’ (Figure 8) Despite a large number of positive words, tweets showed disgust and anger, while fear also make a lot of appearances. The word clouds have shown that the NRC sentiment analysis has been able to identify and categorize some words. For example, the word ‘smell,’ ‘mad,’ ‘hell,’ are in the ‘disgust’ category. Unfortunately, terms like ‘1train,’ ‘4train, and ‘nycsubway’ are categorized under ‘fear.’

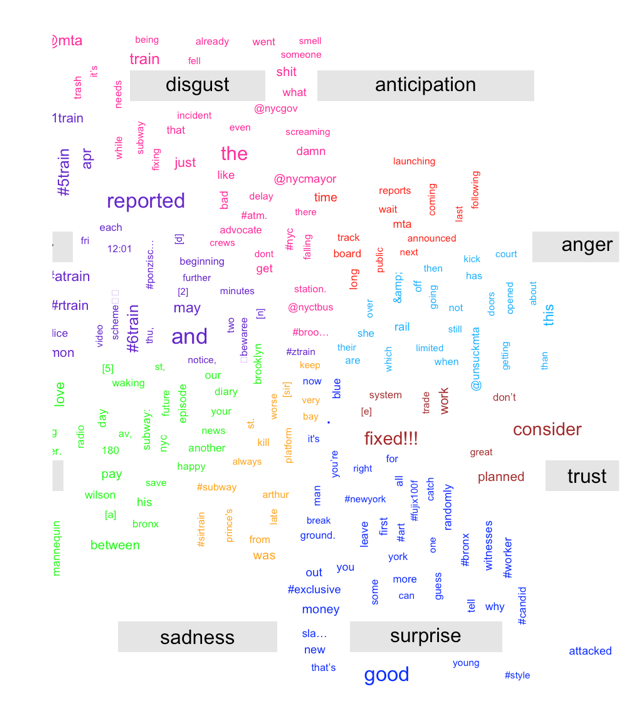
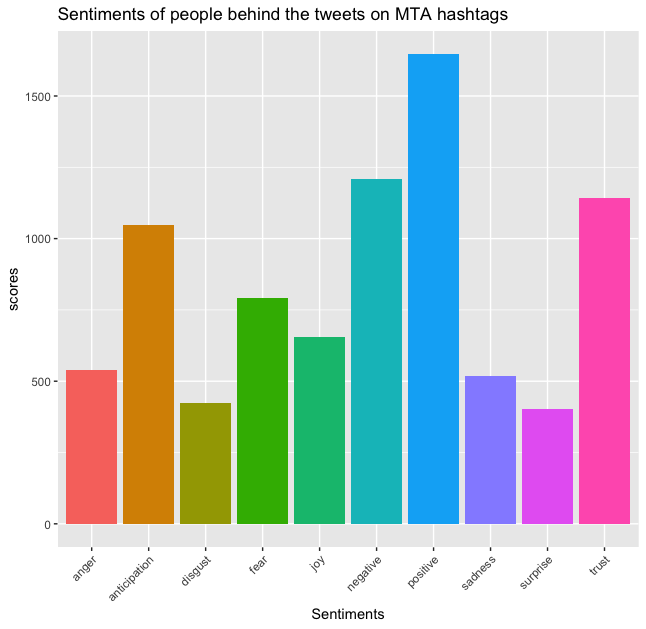


Figure NRC Sentiment Statistics of #mta Tweets

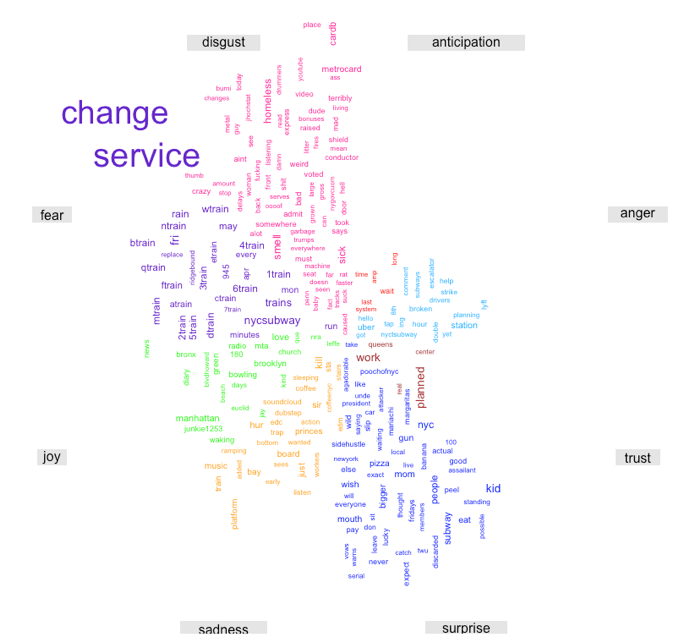
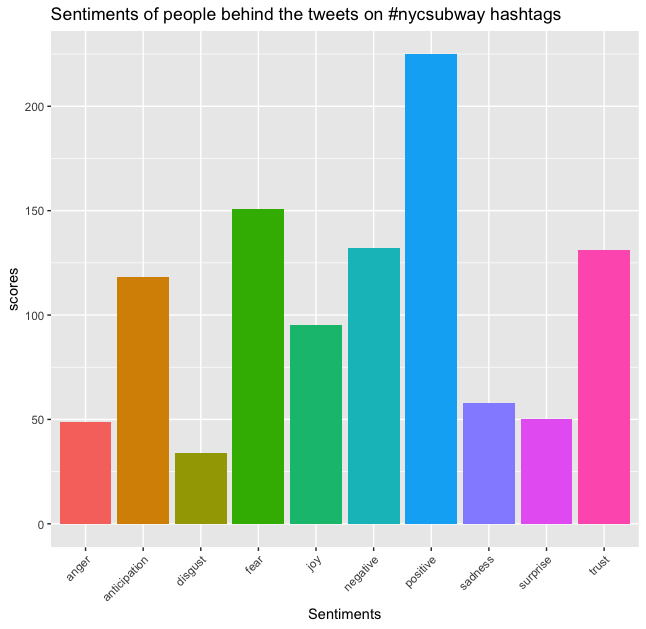


Figure NRC Sentiment Statistics of #nycsubway Tweets

The second part of the analysis focuses on the Emojis as a measure of the contextual sentiment. In order to figure out the emoji with the highest frequency in MTA tweets data, a matrix between MTA tweets and emoji sentiment dictionary was created. The emojis were then ranked by their frequency and top ten emojis with the highest frequencies were shown. (Figure 9)



Figure Top 10 Emojis within MTA Related Tweets

The most common emoji a.k.a. the most popular emoji is the ‘heavy black heart’ within MTA related tweets. It matches the positive and energetic atmosphere among tweets. The ‘Alarm clock’ emoji has a high frequency, which corresponds to MTA tweets’ schedule-related feature. Emoji ‘black sun with rays’ representing sunny days and emoji ‘umbrella with rain drops’ representing rainy days also have high frequency. This shows MTA traffic-related news are also related with weather. (Figure 10)

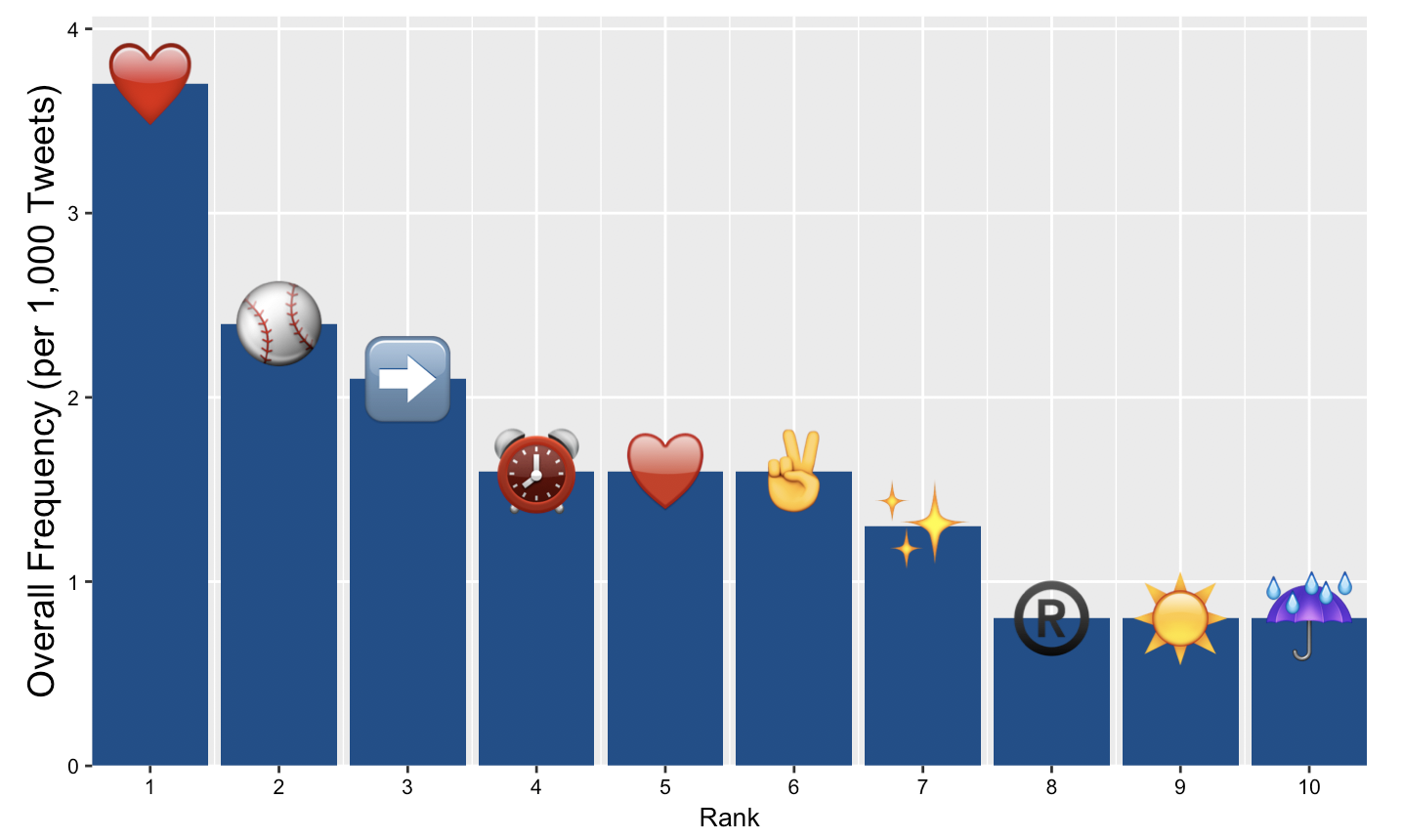


Figure Bar Plot of Top 10 Emojis within MTA Related Tweets

The third part of the analysis seeks to identify subway stations and subway lines with the service updates tweets obtained from ‘@NYCTSubway’ timeline. The ‘@NYCTSubway’ timeline posts real-time updates about subway service changes such as delays and infrastructure malfunctions including signal, switch, mechanical problem, and door problems. The figure below shows a sample service updates tweets posted by ‘@NYCTSubway’ Twitter account. (Figure 11)

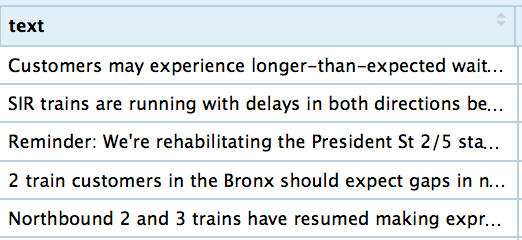


Figure Sample @NYCTSubway Tweets

The top bigrams show some delay reasons that appear quite frequently. The top one is ‘proceeding delays’ with the word ‘with’ removed since the stop words were removed prior to the bigram analysis in the preprocessing. ‘Train service,’ ‘brakes activated,’ ‘running express,’ and ‘medical assistance,’ took place quite often as well, suggesting that they are the top events in the recent subway operation. (Figure 12) After identifying the general causes of delays such as the ‘train brakes were automatically activated’ or ‘someone needs medical assistance,’ the analysis digs deeper to find out the answers to the questions like what stations have delayed most and what lines experienced mechanical issues and what stations need medical assistance most often between April 14th and May 10th. In this way, valuable resources such as emergency responders including EMS, FDNY, and NYPD can be better allocated to combat emergencies and meet customers’ needs. The *rowSums()* and *str\_count()* functions were able to match the station names with the stations appeared in the service updates and count the frequency. As for the result, there was a total of 473 delay-related tweets in the time period. F train, D train, 4 train, and 2 train seem to experienced delay most while the most delay-prone stations include Jay Street Metrotech, Broad Street, 96 Street, 125 Street, and 25 Street. (Figure 13)

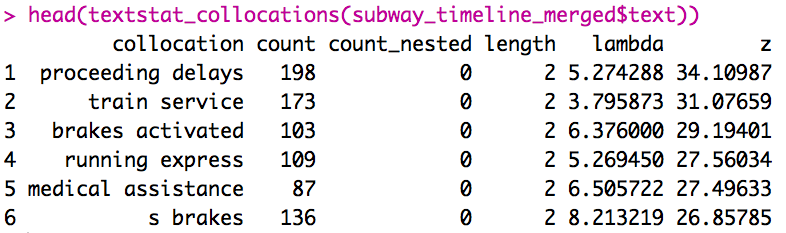


Figure Subway Timeline Bigrams

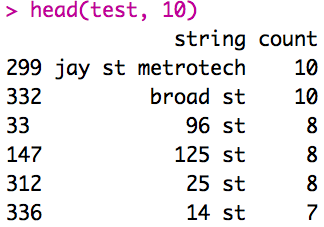
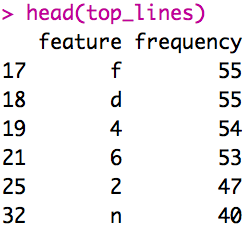


Figure Delay-Prone Lines and Stations

The 7 train Flushing line followed by F train and 4 train experienced more switch malfunctions compared to other lines. The Jay Street Metrotech station yet again makes the top of the list that has experienced the switch problems most. (Figure 14) The station list could provide extra accuracy to the analysis since the Atlantic Avenue station is serviced by F train and 4 train, while the Flushing-Main Street station is serviced by 7 train, etc.

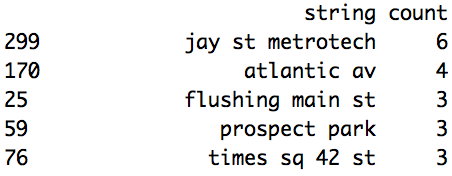
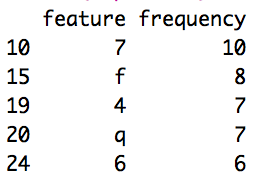


Figure Lines and Stations Prone to Switch Malfunction Issue

As for the inquiries on tweets that contain ‘NYPD investigation,’ the analysis showed that ‘nypd’ appeared 174 in Tweets in the past three weeks. 125 Street station, 25 Street station, and 14 Street station are the top stations experienced NYPD investigations, suggesting that these stations should be stationed with more police force. (Figure 15) What’s more, the 14 Street station, Broad Street station, and 145 Street station received lots of attention from ‘ems’ and medical assistance in the past three weeks.

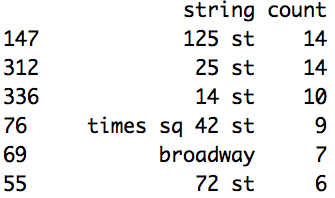
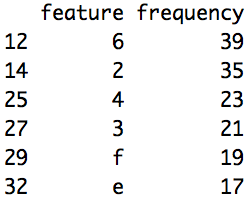


Figure Lines and Stations Prone to NYPD Investigations

# Results Summary

Overall, this analysis has been able to make several observations and identified hotspot subway stations that are prone to service interruptions. Firstly, the Twitter platform is a great source for disseminating transit-related news. However, Twitter users do not engage much with these messages, as the ‘@NYCTSubway’ receive 3 favorite hits and 3 retweet counts per tweet on average despite having 977,000 followers. Secondly, the sentiment analysis showed a generally negative view on transit-related hashtags, yet it failed to reveal specific reasons about subway operation. It also showed a large number of terms in negative moods related to anger, fear, and disgust, although it did not do a good job in classifying certain terms. The last part of the analysis was able to successfully identify hotspot stations that are prone to emergency services and infrastructure malfunctions in a text mining manner. Infrastructures such as certain switches and signals would be able to receive extra attention and urgent repair to prevent future breakdowns. Getting the right fix to the right spot in a timely manner would save commuters tons of money and time, and prevent larger scale service interruptions since multiple subway lines often share tracks and stations across the MTA system.

# Discussion of Limitation and Future Work

Despite the relatively good performance of the analysis, the analysis can be further improved in several ways. Firstly, the analysis needs more dataset to be more accurate statistically. Due to the limitation of the Twitter API and time frame, only three weeks of data were accessed. Secondly, subway line related issues in terms of key words can be analyzed by going through hashtags of specific train lines such as ‘#btrain,’ ‘dtrain,’ and ‘7train’ in the future work. The existing analysis was not able to find line specific issues such as sanitary conditions, delays, medical assistance, or homeless conditions. Thirdly, more hashtags related to a specific train route or bus route does not necessarily suggest that a particular route is more prone to accidents, delays, or interruptions. There are several reasons contribute to this phenomenon. For example, the number of passengers that is taking this route is greater than that of other routes. Thus, even the performance measure for this route is no different from other routes in terms of on-time arrival rate, extra wait time, or cleanliness, there is likely more twitter post about this route. Thus, a weighted ratio should be taken into consideration by comparing with the number of passengers who take a particular route. This adjusted ratio would be relatively more accurate in describing an incidence. Lastly, the emoji sentiment score analysis on tweet data by R is not accurate enough so far. When extracting emojis from tweets, most of the researchers ignore a common situation. The emojis in a large number of tweets are not separated with text data by blank space or punctuations, which makes it complex to match emoticons in tweets. There isn’t much resource available for this particular issue. Most research dropped this kind of tweets from their data when they have a large dataset consisting of 30,000 tweets. However, our dataset of merely 4,000 lines tweets is relatively small compared to it. In the future, we plan to figure out a solution to extract emojis from larger dataset even they are embedded in texts to improve the analysis accuracy.

# Bibliography

Edison Research; Triton Digital; Salesforce.com. (2019). *Percentage of U.S. population with a social media profile from 2008 to 2019* . Retrieved from Statista: https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/

Collins, C., Hasan, S., & Satish V. Ukkusuri, P. (2013). A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data. *Journal of Public Transportation*.

Cottrill, C., Gault, P., Yeboah, G., Nelson, J. D., Anable, J., & Budd, T. (2017 ). Tweeting Transit: An examination of social media strategies for transport information management during a large event. *Transportation Research Part C*, 421-432.

Desai, M. (2018, May). *Sentiment Analysis using R and Twitter*. Retrieved from TabVisExplorer: https://tabvizexplorer.com/sentiment-analysis-using-r-and-twitter/

Dimitrios, E., & Constantinos, A. (2012). Use of social media for transport data collection. *Procedia - Social and Behavioral Sciences*, 775-785.

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015, September 14). *Emoji Sentiment Ranking 1.0*. Retrieved from Clarin.si Repository: https://www.clarin.si/repository/xmlui/handle/11356/1048#

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Liu, J. H., Shi, W., Elrahman, O. (., Ban, X. (., & Reilly, J. M. (2016). Understanding social media program usage in public transit agencies. *International Journal of Transportation Science and Technology*, 83-92.

MTA. (2012). *2012 Customer Satisfaction Survey: Local Bus.* New York City.

MTA. (2018). *Subway Performance Dashboard*. Retrieved from http://dashboard.mta.info/

Peterka-Bonetta, J. (2015, October 16). *Emoticons decoder for social media sentiment analysis in R*. Retrieved from Opiate for the Masses: http://opiateforthemass.es/articles/emoticons-in-R/

Ruiza, T., Marsb, L., & Sernac, R. A. (2016). Social Networks, Big Data and Transport Planning. *Transportation Research Procedia* , 446-452.