Predicting Transit Ridership with Public Perception (using Twitter data)

Miaoyan Zhang 32499705
CPLN 550 Final Paper
Dec. 16, 2019

1 Introduction

Predicting the ridership of any transportation facilities is very important, especially for transit. Because transit ridership is an important basis for urban transportation planning, construction, and operation. The accurate prediction helps to develop better route planning, vehicle selection, and scheduling design. The public perception may be one of the most important factors of the prediction of transit ridership. For instance, if the public generally feels very satisfied with one metro system, the ridership may increase in the following years, and those metro lines that people complain about dirty and messy may have fewer and fewer passengers. The reviews and comments about certain transit systems on the internet may be one of the best ways to get the public attitude towards the system. Those comments with positive sentiment may represent the public's preference for taking this transit system, and those comments with negative attitudes show the public's unwillingness.

This paper is going to crawl transit-related Twitter data, analyze the public attitude from it, explore the trend of transit ridership, find the relationship between public perception and ridership to see whether predicting ridership based on public perception is feasible or not.

2 Methods

2.1 Data wrangling

This paper uses two main data sources. The transit ridership data and public attitude data

The ridership data is download from The National Transit Database (NTD). This data contains all Monthly module data reported to the NTD from January 2002 to October 2019, which is the latest and longest-lasting database we can find. It is reported by mode and type of service. It includes unlinked passenger trips, vehicle revenue miles, vehicle revenue hours and vehicles operated in maximum service (peak vehicles). This paper uses unlinked passenger trips as the ridership of each transit system. The raw data format is shown below.

5 digit NTD ID	4 digit NTD ID	Agency	Active	Reporter Type	UZA	UZA Name	Modes	TOS	JAN02	FEB02	MAR02	APR02
00018	0018	Ben Franklin T	Active	Full Reporter	171	Kennewick-Pasco, WA	VP	DO	49,314	48,281	49,352	48,759
00019	0019	Intercity Transi	Active	Full Reporter	195	Olympia-Lacey, WA	CB	DO				
00019	0019	Intercity Transi	Active	Full Reporter	195	Olympia-Lacey, WA	DR	DO	10,133	9,413	10,229	10,070
00019	0019	Intercity Transi	Active	Full Reporter	195	Olympia-Lacey, WA	MB	DO	216,402	209,621	208,813	224,281
00019	0019	Intercity Transi	Active	Full Reporter	195	Olympia-Lacey, WA	VP	DO	21,286	18,892	19,676	20,606
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	DR	DO	24,311	23,639	26,584	24,534
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	DT	PT				
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	FB	DO				
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	FB	PT	23,932	21,628	24,301	26,055
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	MB	DO	312,681	293,183	320,852	331,235
00020	0020	Kitsap Transit	Active	Full Reporter	180	Bremerton, WA	VP	DO	15,185	13,906	14,938	15,751
00021	0021	Whatcom Tran	Active	Full Reporter	275	Bellingham, WA	DR	DO	12,900	12,284	12,109	13,422
00021	0021	Whatcom Tran	Active	Full Reporter	275	Bellingham, WA	DR	PT	160	238	253	265

Figure 1 Raw ridership data

In order to make the data more readable for further ridership exploration, missing values are replaced by 0. Form the data we can see that each agency contains many modes, this will cause problems when crawling public comments, because people comment on certain agencies like “SEPTA” is much more than on a certain mode like “Philly Heavy rail and Trolley”. Thus, sum of ridership given by each agency each month is counted. Then, the 'melt' function in python is applied to convert the data into long-form and Time-Series by Mode. Text processing is applied then to change the date column to timestamp format for python to read in time format. Because the number of days in each month is different, it may cause deviations in the following analysis. The "daysinmonth" function in python is applied to get the number of days in each month, and daily average ridership of each agency is calculated.

The public attitude data is requested from Twitter developer API. The Twitter API rate limit window duration is 15 minutes. So, each call of search word needs a 15-minute interval. And for the ‘standard’ account, only latest 10 days’ data is available. For each transit agency, a matching search term is created, such as " Southeastern Pennsylvania Transportation Authority" corresponding to "SEPTA". Use the search term to request relevant Twitter data through the API and 15 minutes between each request. Due these reasons, this paper only searched the Twitter data corresponding to the 15 largest agencies in the previous step.

2.2 Text mining

In this paper, NLKT package and TextBlob package are used to process textual data.

Using NLKT package, what is the most common word people mention about the transit system is going to be explored. Firstly, all URLs are removed to get the regular expressions. Then all common words that do not carry too much significance, and which are often ignored in many text analysis (stop words) are also removed. Thirdly, the query terms which have just used to left only the meaningful word are removed. With these words, we can get a general idea of the attitude of the public toward a certain transit system.

With the TextBlob, we are going to make sentiment analysis, to find out the attitude or emotional of the person who sent the particular tweet about the certain transit system. The TextBlob package uses pre-trained machine learning algorithms to classify whether the text is positive or negative and whether it is subjective or objective. The polarity is represented with a float number within the range from -1 to 1 where -1 is negative and 1.0 is positive. The subjectivity is represented with a float number within the range from 0 to 1 where 0 is objective

and 1.0 is subjective.

2.3 Exploration and Relationship

It is very meaningful to predict ridership through public perception. However, due to Twitter API permissions, we can only get data for the last 10 days, not as long as the ridership data. Besides, because the total number of requests is limited, we can only get the public perception of the Top 15 ridership transit systems. This data is not enough to build a machine learning model, but under the existing data conditions, we are able to explore the trend of ridership change and the difference of public perception on different transit systems, and Pearson correlation coefficient is calculated to analyze the relationship between ridership and public perception. The Pearson correlation is shown below:

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = r$$

The Pearson correlation help to measure the linear correlation between the public perception, which is the polarity and subjectivity, and the transit ridership. Pearson correlation is used in this step is because it does not depend on units of measurement of perception and ridership.

If higher-level permissions is available, a time related machine learning model can be built to get a more accurate prediction.

3 Findings

3.1 Ridership exploration

After data preprocessing mentioned above, the result dataset is shown below. 15 agencies with the highest ridership across the country are selected. Ridership data start from January 2002 to October 2019. We have the unlinked passenger trips of each month and trips per day of this month.

	Agency	Date	UPT	year	month	Date_1	trips	day_in_month	trips_per_day
0	MTA New York City Transit	2-Jan	225702964.0	2002	Jan	2002-01-01	225702964	31	7.280741e+06
1	Chicago Transit Authority	2-Jan	39536607.0	2002	Jan	2002-01-01	39536607	31	1.275374e+06
2	Washington Metropolitan Area Transit Authority	2-Jan	31114119.0	2002	Jan	2002-01-01	31114119	31	1.003681e+06
3	Los Angeles County Metropolitan Transportation...	2-Jan	37468791.0	2002	Jan	2002-01-01	37468791	31	1.208671e+06
4	Massachusetts Bay Transportation Authority	2-Jan	28763910.0	2002	Jan	2002-01-01	28763910	31	9.278681e+05
...	...	...	...	...	...	...	...	...	...
3205	San Francisco Bay Area Rapid Transit District	19-Oct	11597346.0	2019	Oct	2019-10-01	11597346	31	3.741079e+05
3206	Metropolitan Atlanta Rapid Transit Authority	19-Oct	10624986.0	2019	Oct	2019-10-01	10624986	31	3.427415e+05
3207	MTA Long Island Rail Road	19-Oct	10221898.0	2019	Oct	2019-10-01	10221898	31	3.297386e+05
3208	Denver Regional Transportation District	19-Oct	9597573.0	2019	Oct	2019-10-01	9597573	31	3.095991e+05
3209	Tri-County Metropolitan Transportation Distric...	19-Oct	8842859.0	2019	Oct	2019-10-01	8842859	31	2.852535e+05

Figure 2 Monthly ridership data

The line chart below visualizes the monthly ridership data. We can see that the ridership of the MTA New York City Transit is much higher than other transit systems, for about 5 times or more. Nearly 10 million trips happen per day in MTA New York City Transit system. We can also see a general growth trend in this system's ridership. Chicago Transit Authority is the second highest, with about 1.5 million trips per day. The MTA bus company started at 2005 and have no trips before 2005.

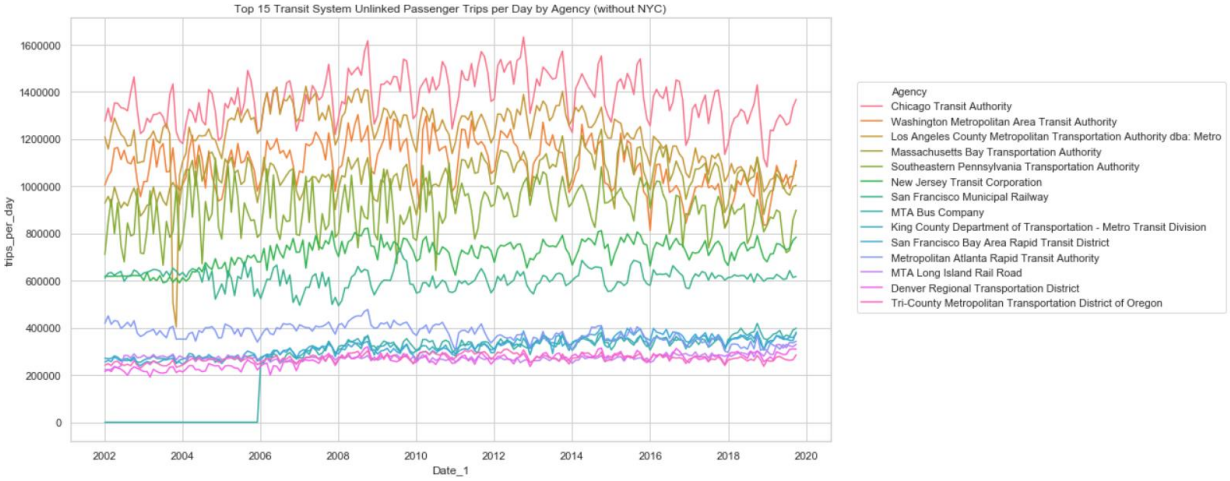

Figure 3 Monthly ridership line chart

Since ridership of NYC MTA is too high, we removed it to see the ridership trends of the remaining 14 transit systems shown in *Figure 4*. We can see that the ridership change happens on an annual basis, the ridership is always very low at the beginning and the end of the year, but high in the middle of the year. Besides, there are always two declines followed by a rebound in the middle of each year. It is worth noting that the trend is almost the same between different agencies each year. In addition, the larger the ridership, the larger the fluctuations of the transit system, and the smaller the ridership, the smaller the fluctuations of the trips. It seems that the proportion of data fluctuations is similar.

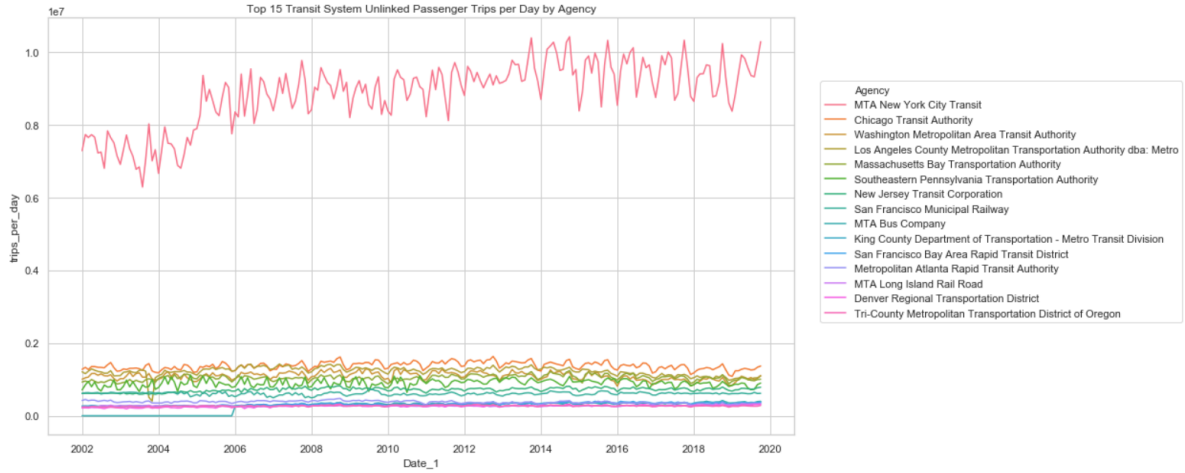

Figure 4 Monthly transit ridership line chart (without NY MTA)

In order to more clearly observe the overall change trend of the data, the total number of transit trips each year is summed to remove the slight fluctuations in each year. The general trend is shown in *Figure 5* below. Because only 10 months is included in 2019, thus there is a decreasing trend present in the chart, which is not the real case. We can see that the overall change trend is not obvious.

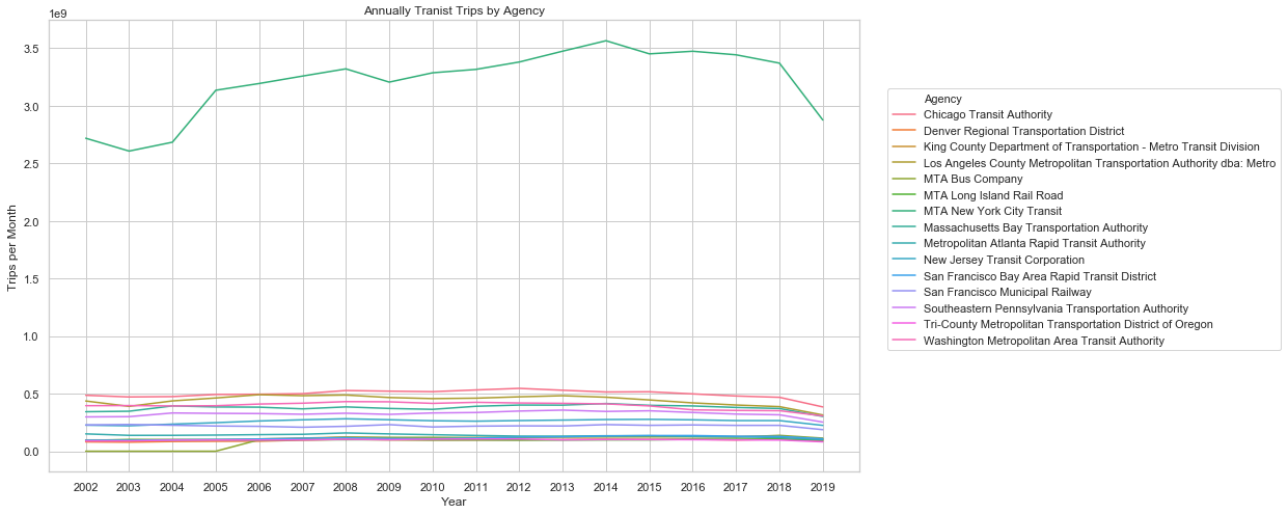

Figure 5 Annually transit ridership line chart

Removing NYC MTA, we get *Figure 6* shows the annually transit ridership of rest agencies. We can see that many transit agencies have shown a downward trend on ridership since 2013. This may be caused by the appeal of Internet companies such as Uber has snatched some of the original transit passengers.

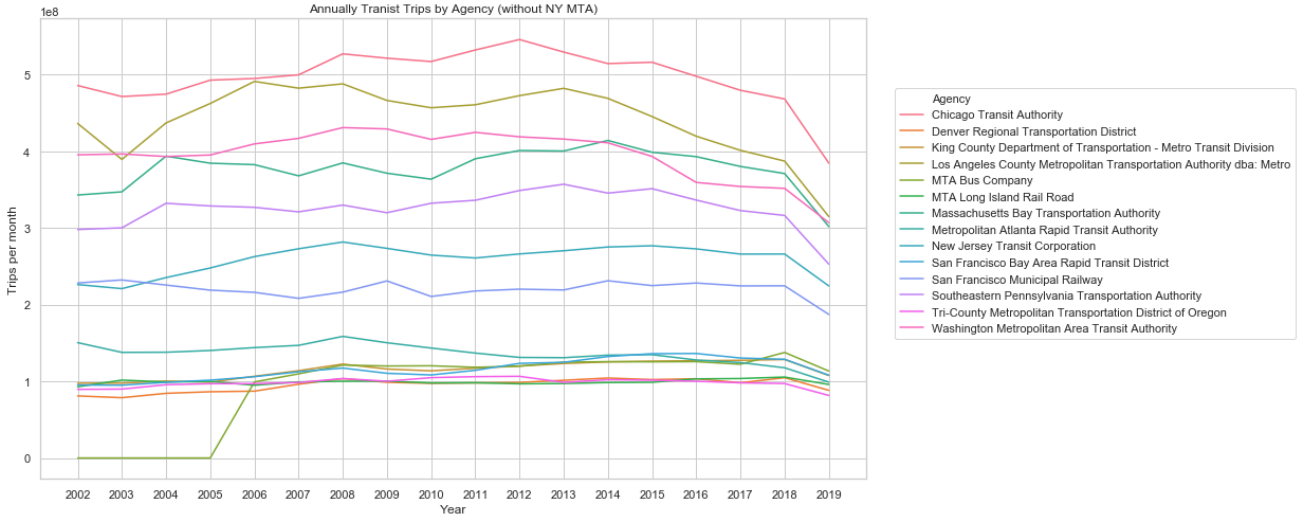

Figure 6 Annually transit ridership line chart

To further explore the trend of ridership change, we calculate the increment in average trips per day between two consecutive months (The difference between two months divided by the previous month). We can see that in most cases the change in trips remains within 20%. The change of SEPTA before 2011 was very unstable, with each increase or decrease exceeding 20% all the time. There is a huge decrease and rebound that happened in 2003 on the Washington Metropolitan Area Transit Authority.

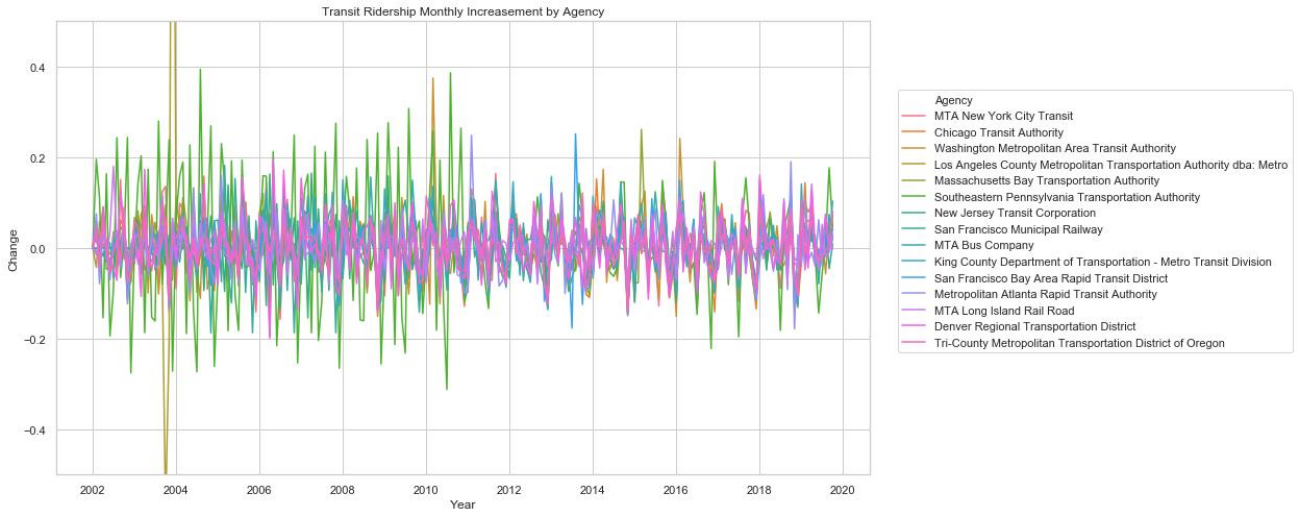

Figure 7 Monthly transit ridership growth rate

The subset of last two years is shown in *Figure 8*. We can see that the growth rate often increase in one month and decrease in the next month, and rarely increases or decreases simultaneously for two months. Except November and December, which is caused by holiday and cold weather. Therefore, if the exploration analysis is on a monthly basis, the data will be greatly affected by fluctuations.

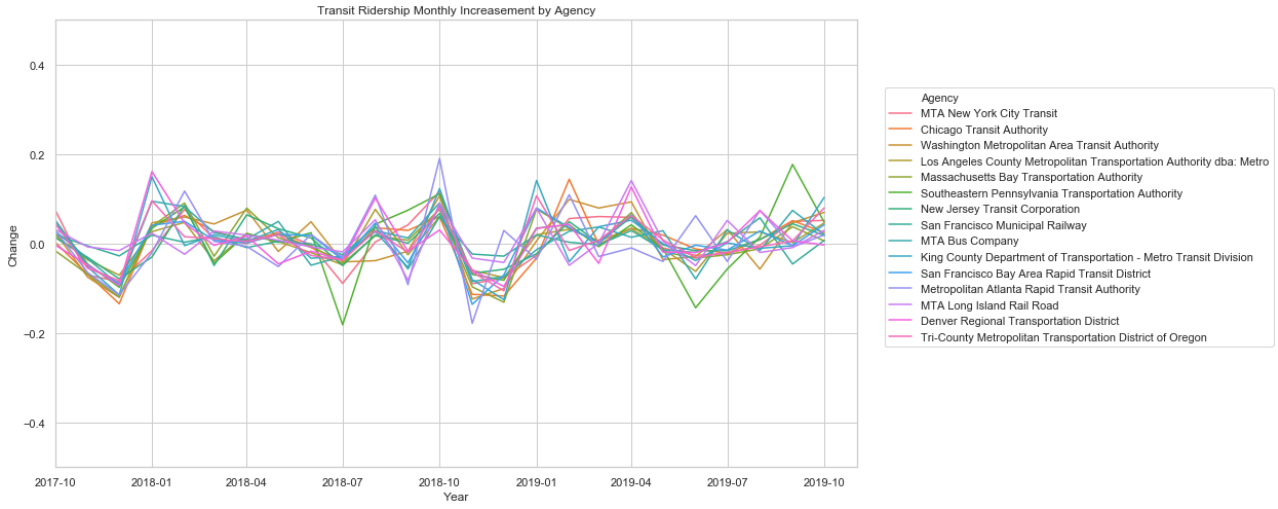

Figure 8 Monthly transit ridership growth rate (recent 2 years)

In order to reduce the impact of the data fluctuations, the annual trips are summed up to calculate the growth rate of trips between two years, as shown in *Figure 9*. With only 10 months of data, the reduction in 2019 is negligible. There are still some fluctuations. From 2015 to 2018, the ridership of most transit agencies is declining. Denver Regional Transportation and MTA bus bottomed out in 2018, and Los Angeles Metro maintained steady growth for a long time.

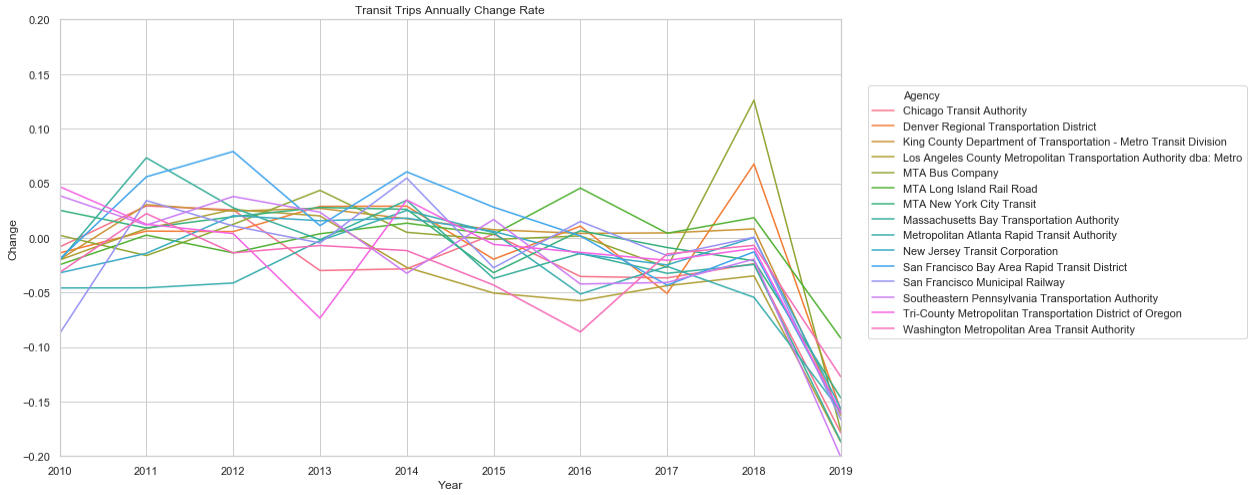

Figure 9 Annually transit ridership growth rate

In the next step, this paper is going to explore public perceptions of these agencies, to see what we can find to help develop a relationship analysis.

3.2 Public perception exploration

After requesting data from Twitter API as described in the method. We get the twits as shown below. It shows the Twitter content results by searching "CTA Chicago". These twitters display how twitters represent the public perception. Some display how users think about the transit service and others reflect user's opinions about some issues, but generally show public attitudes toward the transit service.

Table 1 Twitter example

agency	**text**
CTA Chicago	@nbcchicago People at home in Chicago afraid to go outside to walk, shop and take CTA. We very well might be beaten or killed. The mayor is out of town…
CTA Chicago	Heading to EdgeH2O Beach (at @CTA Bus 36 in Chicago, IL)
CTA Chicago	Veteran urban planner is Lightfoot's choice to serve as transportation commissioner - Chicago Sun-Times: #CTA #Metra #Transit #Chicago #Transportation
CTA Chicago	looking north from the Diversey Brown Line station #2019favoritephotos #chicago #cta
CTA Chicago	RT @cta: Blue Line riders, don't forget: Forest Park trains will bypass Grand starting at 8pm thru 4am Mon 12/7, as station is closed for…

By removing all common words that do not carry too much significance, and which are often ignored in many text analysis, and remove all punctuation, we can calculate the most common words in these Twitters, as shown in *Figure 10.*

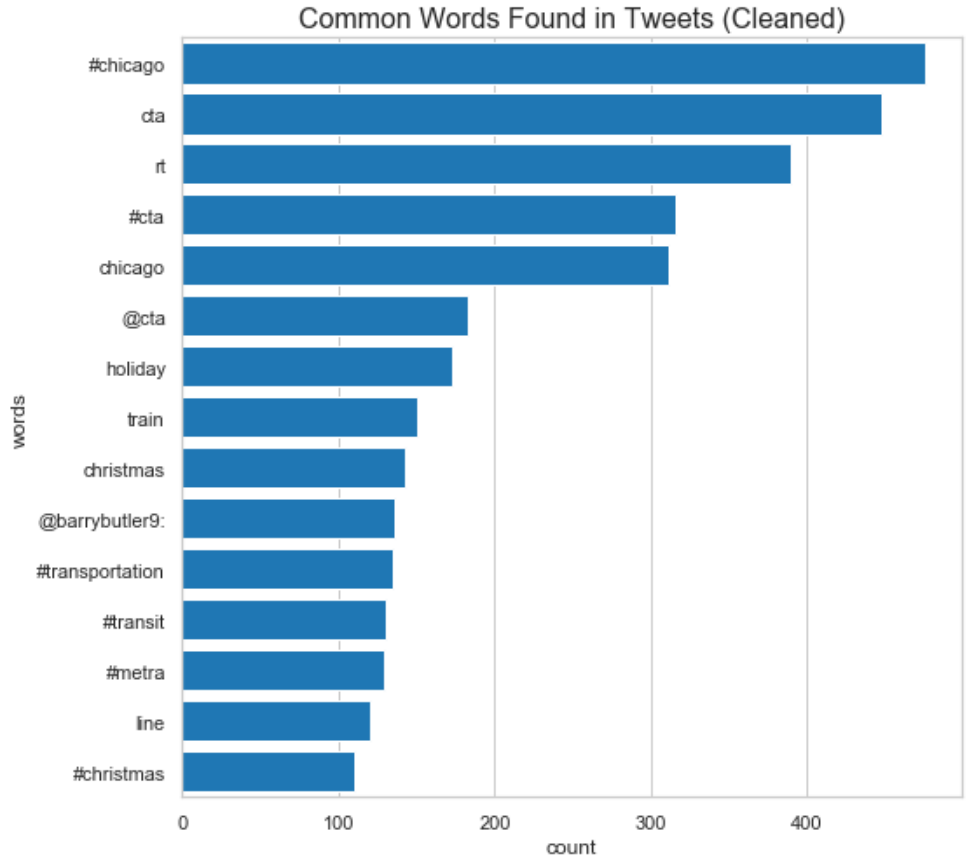

Figure 10 Common words about CTA Chicago

By sending these twitters to TextBlob, we can get the sentiment of each twitter, as shown in *Table 2.*

Table 2 Twitter sentiment example

agency	polarity	subjectivity	text
CTA Chicago	0	0.033333	😖: The overall mood of riders worsened in the ...
CTA Chicago	0.25	0.9	#RSNA19 Tip: Did you now that the train will t...
CTA Chicago	-0.2	0.4	Two Americans Dead, Five Injured in Belize Bus...
CTA Chicago	0.183333	0.522222	😎: The general mood improved in the last hour....
CTA Chicago	0.1	0.3	The bus is 2 minutes early (at @CTA Bus Stop 1...

Using the create time features of these twitters, we can visualize the public perception on CTA Chicago over time, as shown in *Figure 11*. We can see that there is no obvious trend in the data. One possible reason is that Twitter can only provide data for the last 10 days, and the duration of the data is too short to extract features.

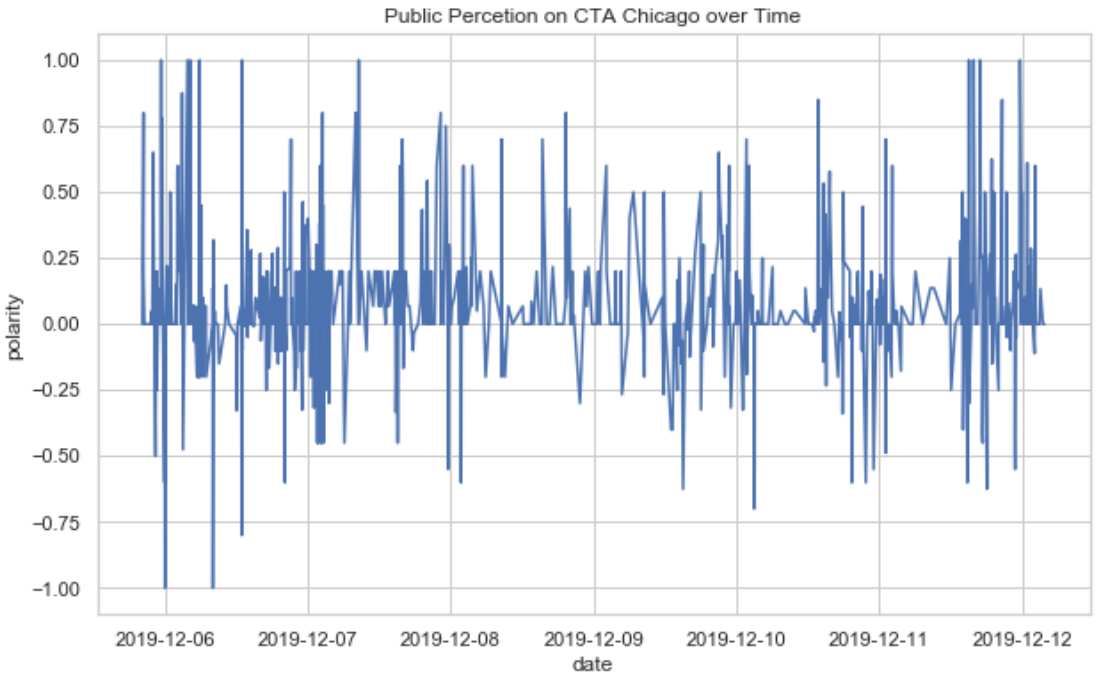

Figure 11 Public perception on CTA Chicago over time

By repeatedly doing the same request and sentiment analysis work to all 15 transit agencies, a dataset with 18971 twitters and their corresponding sentiment is created. The result box plot is shown in *Figure 12* below. We can see that twitters about King County Metro have the most positive attitude and twitters about SEPTA and New Jersey Transit have the most negative attitude. People's attitude towards TRIMET is the most concentrated, neither too negative nor too positive. Generally, for most transit agencies, public's perception is positive. And from the lower chart we can see that for most transit agencies, public's perception is objective. Titters that mentioned BART are the most subjective and Titters that mentioned Los Angeles Metro are the most objective.

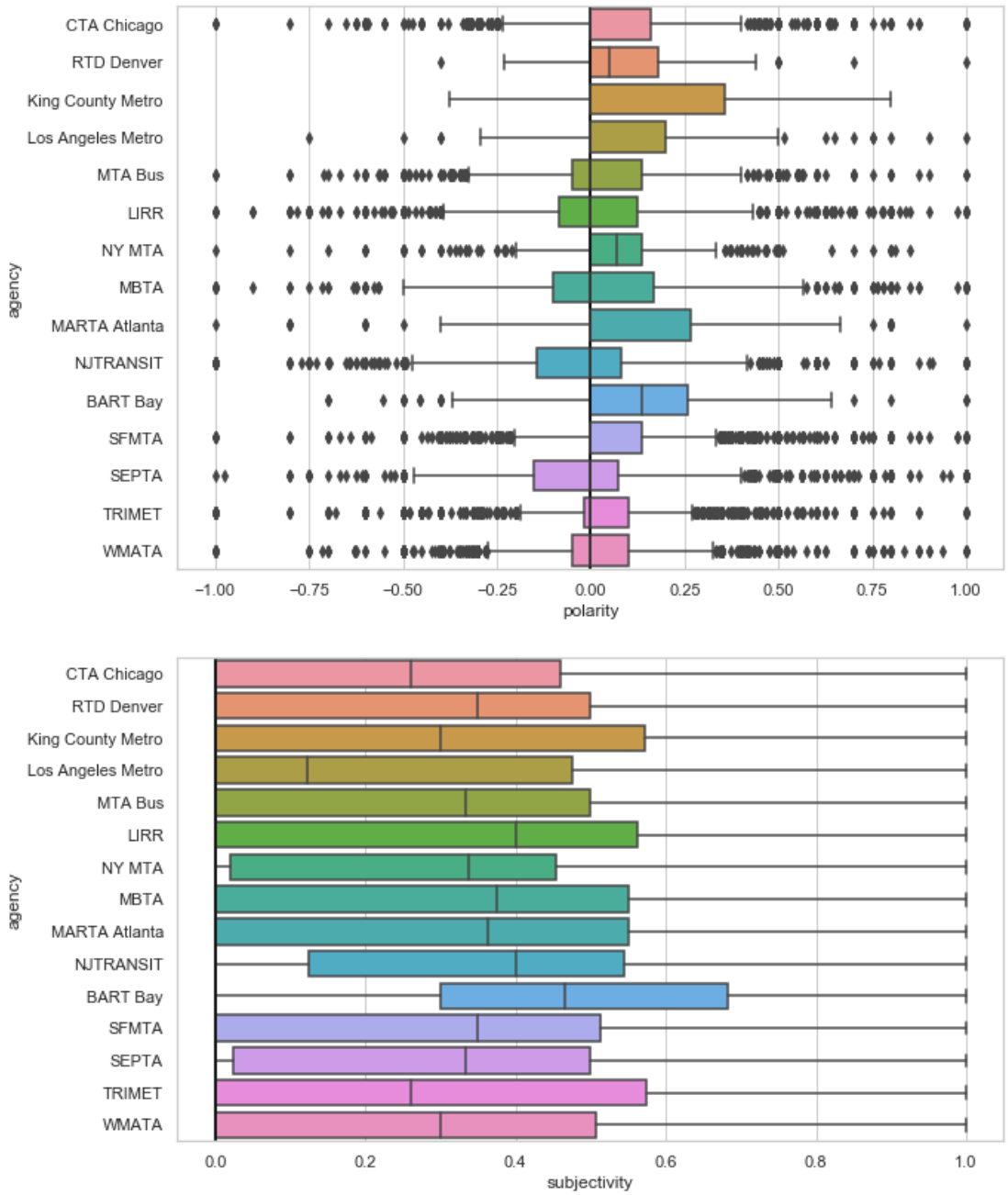

Figure 12 Public perception box plot

3.3 Relationship between ridership and public perception

Through the first two sessions, we explore the trend of transit ridership change and the emotional characteristics in twitter. In this session, we are going to further explore the relationship between these two features.

Figure 13 shows the relationship between passenger trips of 15 different transit agencies in 2019 and public perception polarity about these agencies. The ridership of MTA New York City Transit is much higher than other transit systems, which makes the relationship difficult to describe.

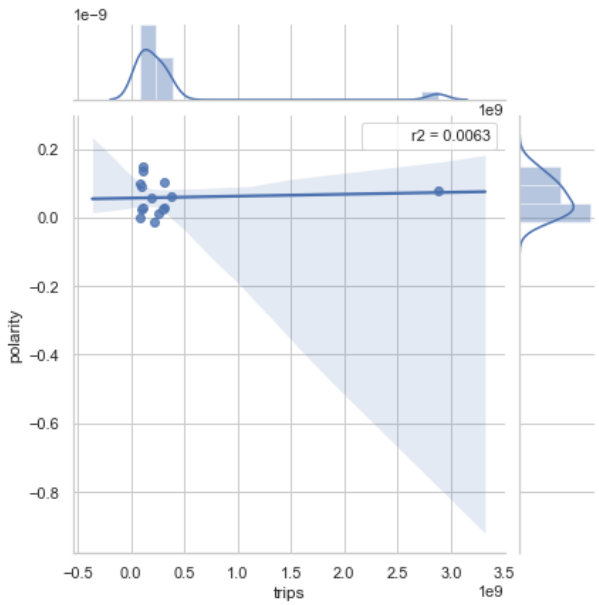

Figure 13 Correlation between transit trips and public perception polarity

Removing NYC MTA can get *Figure 14*, which shows the relationship between ridership of rest agencies and public perceptions. We can see that as the number of passengers in the transit system increases, the public's perception gradually becomes negative. But the correction between two factors is only 0.05, which is very weak. Thus, we can't say there is a relationship between trips and polarity. Using the polarities of public sentiment to predict ridership may not be useful.

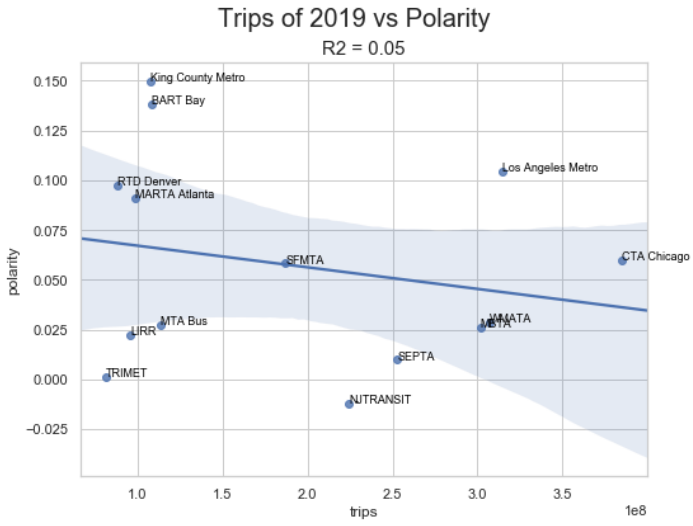

Figure 14 Correlation between transit trips and polarity (without NY MTA)

Then, let's see the relationship between trips and twitter's subjectivity. We can find out that people are more objective on the transit systems with higher passenger trips. However, the Pearson Correlation between two factors is 0.12, still not substantial.

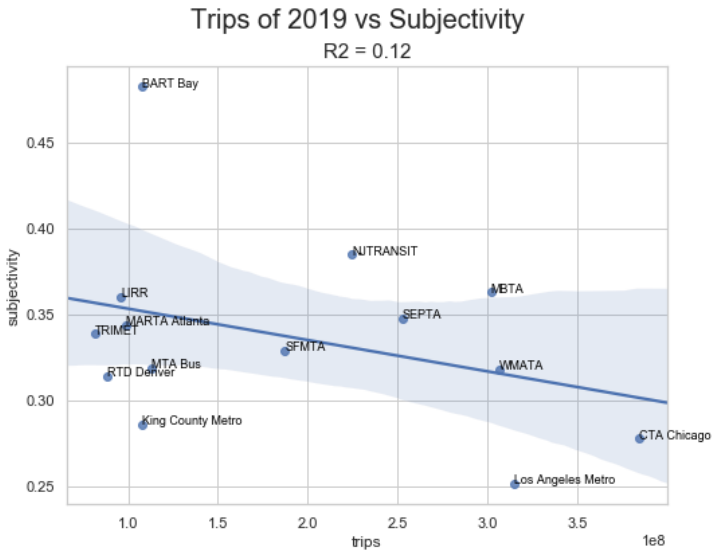

Figure 15 Correlation between transit trips and subjectivity (without NY MTA)

Using public perception to predict transit ridership may not be a good idea because it is more affected by the size of the transit system rather than by whether people are willing to ride. In contrast, it may be more feasible to use public perception to predict the increase and decrease of transit ridership. Comparing the changes of a single agency can avoid the prediction error caused by the different scales of different agencies.

In the previous exploration of the transit ridership, we find out that the change of ridership takes an annual cycle, and the change trend in each year of different agencies are about the same. In order to get rid of these recurring changes in cycle form and to accurately describe the impact of public perception, the trips over the past two years are calculated (November 2017 to October 2018 and November 2018 to October 2019). Sum them separately to calculate the growth rate and then performed correlation analysis in public perception. The result is shown in *Figure 16* and *Figure 17* below.

From the result we can see that, people with negative attitude is likely to increase the passenger trips, which is contrary to our hypothesis that positive emotions should lead to increased ridership. Besides, neither the polarity nor the subjectivity shows a strong correlation with the transit passengers ridership, especially the Pearson Correlation of subjectivity and increment is only 0.01.

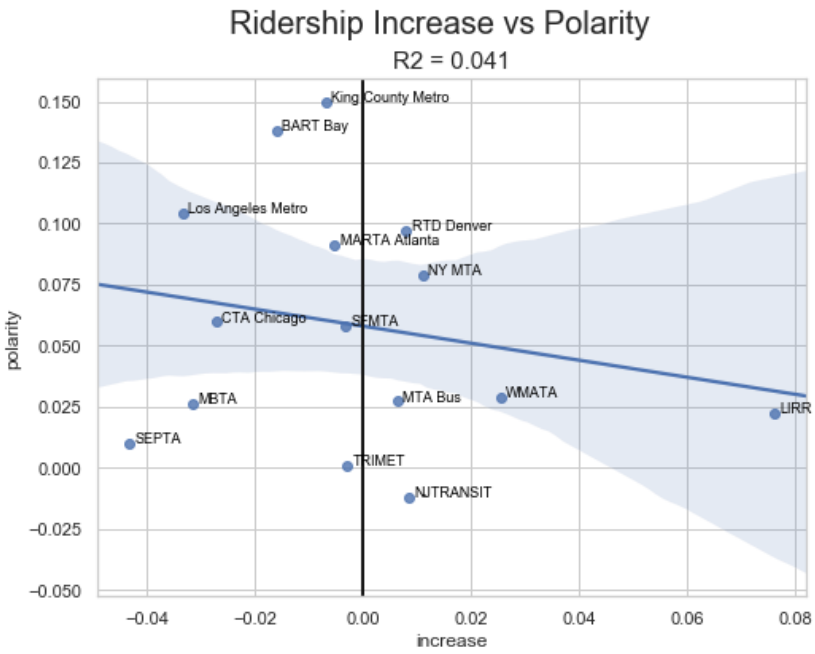

Figure 16 Correlation between transit trips increase and polarity

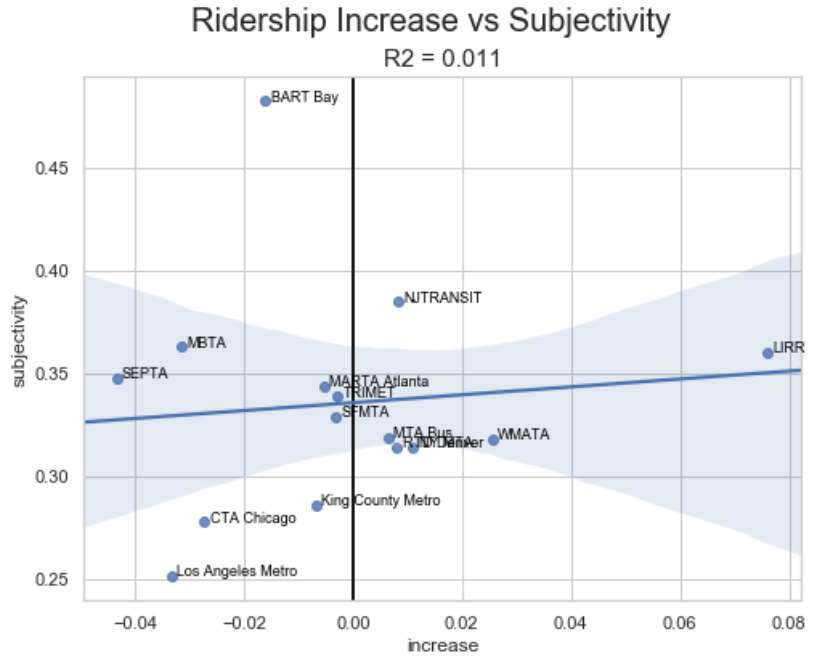

Figure 17 Correlation between transit trips increase and Subjectivity

So far, we can conclude that with the existing data we cannot find a correlation between public perception and transit ridership, so we cannot predict transit ridership through public perception. There are many reasons for it.

Firstly, incomplete data is an important reason. Because of the limited permissions of the Twitter API, only a small amount of transit agencies' public attitude data for only the last 10 days is available, which made the correlation analysis insufficient and not enough for reference. Secondly, the relationship

between public attitudes and the transit ridership is complicated. The public's preference for public transportation may indeed cause more people to travel by transit, but the increase in the number of passengers in transit will also lead to a decline in the ride experience and therefore lead to more negative emotions. This interaction may be difficult to describe through data.

4 Meanings

Based on the above research, this paper found that it is impossible to predict the transit ridership using the existing public perception data, but if there is more public attitude data, this prediction is likely to be completed.

In subsequent studies, longer-term public perception data can be used. Using these data, you can explore the changing trend of public perception, discover its changing rules, and compare it with the changes in transit ridership to find the similarities, differences, and mutual relationships between the two. The use of public attitude changes may collectively complete the forecast for transit ridership, because public attitude changes that are more positive may be more likely to cause transit ridership to rise than public attitude positives and vice versa. At the same time, this research method can also be applied to the forecasting of ridership at different stations of a transit line. Due to the same number of vehicles operating at different stations on the same bus line, public attitudes may have greater impact on ridership.

In addition, this article summarizes many features of transit ridership and public perception. For example, the trend of ridership over time, the comparison of ridership and public perception between different transit agencies, and so on. These studies are valuable for future research.

5 Conclusion

First, this paper explores the changes in transit ridership access in the last 17 years. The research object is the 15 agencies with the highest ridership, and it is found that the number of passengers in the New York MTA is much higher than other public transportation systems. The ridership changes on annually basis, and this trend is similar in all agencies. Besides, there has been a slight downward trend in transit ridership in the last 6 years. Secondly, this paper used the Twitter developer API to crawl the above 15 public transit agencies-related tweets. And the semantic analysis has been done on these tweets from polarity and subjectivity. Finding out that the public's attitude towards transit was overall positive and objective, but the attitudes towards different transit agencies were slightly different. Third, this paper conducted a study on the relationship between transit

ridership and public perception, and found that the correlation between the two is very weak, both in terms of the total number of passengers and the increase in the number of passengers. The underlying interactive causal relationship between the public perception and ridership, and the incompleteness of the data may be to blame.

In addition, there are still many shortcomings in this paper. Because the API data can only be extracted once every 15 minutes for only the last 10 days, and the total number is limited. This study lacks extensiveness, and the trend of public perception has not been extracted. For the existing public attitude data, only the analysis of polarity and subjectivity is performed. It may be better to extract more emotions, such as anger, happiness, satisfaction, loss, etc. These tones can also be used as predictors of ridership.