Relationship among Crimes, Public Commuters, and Sentiment in Manhattan

Zhicheng Ding
Computer Science
Columbia University
Email: zd2212@columbia.edu

Yunan Lu
Electrical Engineering
Columbia University
Email: yl4021@columbia.edu

Lin Bai
Electrical Engineering
Columbia University
Email: lb3161@columbia.edu

***Abstract*—A great amount of data in many different fields have been collected in the recent year. These high volumes of data provide a great chance for us to explore the relationship among data which seems vogue in the previous time. In this project, we explore the relationship among three fields of data (*i.e.* crimes, public commute, and social network). We first collected the crime, subway, taxis, and twitter datasets with the timestamp and geographical location in the year 2014. Then we aggregated the data by spatial and temporal information in Manhattan to show the trend hourly by hourly in an average week. In addition, we designed three different charts for using Vue.js and D3.js to explore the data. Lastly, the result demonstrates that strong positive linear correlation of the number of crimes and the number of commuters in most the areas in Manhattan. The average sentiment of Manhattan also decreases (*i.e.* less positive) during workhour and increase (*i.e.* more positive) after work and during the week.**

***Keywords*—Big Data Analysis, Vue.js, D3.js, Crime Map, Public Commute, Sentiment Analysis, Nature Language Processing.**

I. INTRODUCTION

In the era of the digital world, data are collected in various ways. High volumes of data provide us with some breakthroughs in various fields when their trends in vogue in the previous time. The big data analytic give an innovative way to think about the problems which are hard to analyze via traditional way.

Crime is a severe threat to society and grows to involve more wide and complex criminal activities with technology help. This makes it even harder to investigate and prevent crime by using traditional ways. Since information on suspects can be geographically diffuse and happened occasionally, data analytics provides an innovated method to uncover hidden information.

Transportation is an important topic for the government. Because of this, public transportation study has emerged and become a significant portion of work for transportation system planning programs in terms of both cost and personnel. And this study is supported by big data collection process and data mining analytics.

From the description above, it seems like that crime and public commute seems to have a similar pattern. So in this project, we aim to find out the pattern, meanwhile, we expected sentiment also follow similar trend. Therefore, we try to explore the relationship among crime, public commutes, and sentiment in Manhattan. We found and download the associated datasets (*i.e.* the criminal dataset in the Manhattan at the year 2014, the subway and taxis dataset of New York City at the year 2014, and the twitter dataset at the year 2014).

To better evaluate the relationship among these data, we designed three different charts using Vue.js and D3.js to visualize the data hour by hour in an average week: (1) a heat map chart of crime and public commutes in Manhattan. This plot provides spatial and temporal information about crime and commutes; (2) a bubble chart of crime, public commutes aggregated by precinct[1] in Manhattan. This plot shows a specific linear correlation of the data; (3) some statistic charts to show the trend of the crime, commutes, sentiment, and their correlation in each day. This plot provided a more clear result of how this relationship changed on a given day. In the web UI, we also provide a story mode which gives a chance for audiences to explore the relationship which some timestamp highlighted. The novelty of this project could be summarized as below.

- We conducted the relationship among the number of crime, number of commuters, and the sentiment of Manhattan people hour by hour in an average week.
- We design three different charts (heat map chart, bubble chart, and statistic chart) to demonstrate the relationship behind the data.
- We extensively create a story mode which highlights the critical date time in this project which provides instruction for users to explore the data.

II. RELATED WORKS

A. Big Data Analysis

Big Data has become a heated topic in recent year with the benefit of significant improvement in computing capability [1]. Recently, big data has shown high potential in both industries and academic [2]. As a result, big data application has developed in many different fields, including: text data analysis [3]–[5], web data analysis [6], [7], social network data analysis [8]–[11] and mobile data analysis [12]–[14].

B. Crime Data Analysis

Previous crime data mining techniques include entity extraction, association analysis, classification and social network analysis used to identify crime patterns in structured and unstructured data [15]. For different crime type, different analysis tools can be applied. Some crime forecasting models

[1]Precinct: a district of a city or town as defined for police purpose.

also developed to identify the potential crime hot-spots and the trend of crimes in specific area [16]. There are also some related work has been done in Kaggle, such as competition: San Francisco Crime Classification, which also reflects the data analytic has been widely accepted and applied on the crime related problem[2].

C. Public Commute Data Analysis

Because of the wealth of information that can be obtained from taxi data, several efforts have focused on their analysis. Xi *et al.* [17] collect taxi trip data from Shanghai to study the travel patterns and city structure in Shanghai which provide the reference for future urban and transportation planning. Veloso *et al.* [18] utilize these data to study human mobility in Lisbon. They explore the patterns of pickup and dropoff location distributions. In addition, Yuan *et al.* [19] study the resilience of taxi and subway trips in New York City to explore post-hurricane recovery patterns of the roadway and subway systems. Some projects also use MTA subway data to classify location according to the usage of land[3]. Besides, due to the importance of this information and the related problem associated with this information, some visualization systems have developed to facilitate related study [20].

D. Twitter Data Analysis

Sentiment analysis is useful to collect and find out what other people think. There are lots of related work have been done on Twitter sentiment analysis and normally divided into two parts: identifying sentiment expressions and determining the polarity of them. The goal is to classify their opinion based on their emotional content as positive, negative and neutral [21]. Some publishers use lexical resources and classified by the presence of lexical items [22]. Some combine conjunction rules as additional features to obtain more accurate result [23]. Barbosa *et al.* [24] further increase the accuracy by considering the re-tweets, hashtags, punctuation and exclamation marks as part of the context to evaluate peoples opinion.

III. SYSTEM OVERVIEW

In this section, we will first show the whole system diagram in Section III-A. Later on, we will describe all the dataset in detailed that we had used in this project in Section III-B.

A. System Diagram

Figure 1 demonstrates the high-level system diagram of our project. We follow the ETL pipeline to get the prepare the data which eventually get loaded and visualized in our website. The homepage of our website is shown as Figure 2. Our project majorly contains three parts:

- **Data collection and pre-processing.** We collect the data of crime (size: 53MB), commutes (size: 26GB), and twitter (size: 313.7GB) datasets. Then, we cleanup and extract the useful features for modeling. We will provide more information regarding data description in Section III-B and data pre-processing in Section IV-A.
- **Data processing.** We utilized the pre-processed data for further processing, including aggregation and modeling, to generate the final data for visualization in our website. To be specific, (1) we aggregated the data by precinct number and coordinate respectively; (2) we fit the data with linear regression and calculate the Pearson correlation; (3) we processed the twitter dataset using NLP based algorithm to analyze the sentiment of Manhattan people. We will discuss it in detail in Section IV-B.
- **Data visualization.** We designed a website using D3.js and vue.js which load the data after data processing. The website contains three different charts: (1) heatmap chart aims to give a spatial and temporal impression to audiences; (2) bubble chart highlights the linear regression of the number of crime and the number of commuters; (3) statistic chart summarizes the data using all data and it demonstrates that the dynamic trend of crime, public commutes, and sentiment. The homepage of this website

[2]https://www.kaggle.com/c/sf-crime

[3]https://www.nycedc.com/blog-entry/using-mta-turnstile-data-examine-land-use

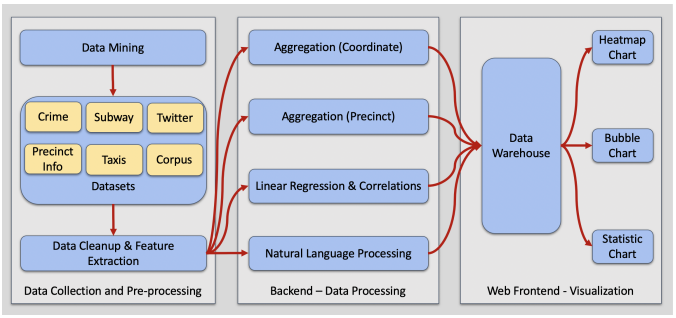

Fig. 1. System overview of project: (1) we firstly collect and process the data (crime, commutes, and twitter datasets); (2) then, we aggregated the data by coordinate and precinct, calculate linear correlation using Pearson correlation, and process context using NLP technique; (3) lastly, we store all processed data and designed web UI using Vue.js and D3.js for visualization.

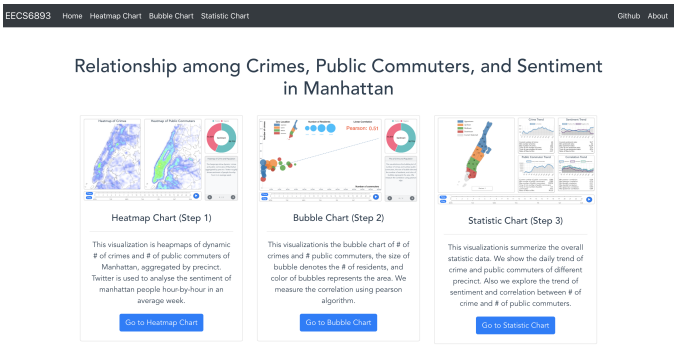

Fig. 2. Homepage of Web UI. We had designed three UIs to explore the data step-by-step: (1) heat map chart shows the spatial and temporal information crimes and commuters distributed in Manhattan; (2) bubble chart demonstrates the positive linear correlation using the Pearson correlation; (3) statistic chart summarizes the overall data trend in one day which helps conduct the relationship among these crimes, public commutes, and sentiment.

is shown in Figure 2. We will discuss it in detail in Section IV-C.

B. Data Description

In this section, we will describe the dataset, including crime, public commutes (MTA subway and yellow taxis), twitter, precinct geojson of New York City, population information of Manhattan neighborhood. We care about the dataset with the timestamp and coordinate information because we can explore the data by spatial (*i.e.* coordinate and precinct number) and temporal (*i.e.* hour and weekday) respectively. With timestamp information, we can convert that into the hour and weekday. With coordinate information, we could calculate if this coordinate inside the polygon, as we already had the polygon information of each precinct. Eventually, we found and downloaded the following datasets, listed below.

1) Crime Dataset: The data of crimes is from the Kaggle: 2014-2015 Crimes reported in all 5 boroughs of New York City[4]. This dataset reports 2014-2015 crimes in all 5 boroughs of New York City, which contains 23 fields. The attributes that we are going to use are id, timestamp, coordinate, precinct number. The total size of this dataset is 53MB. This dataset will further process as two datasets: (1) processed data1 contains weekday, hour, precinct number, and the number of crimes; (2) processed data2 contains weekday, hour, coordinate, and the number of crimes.

2) Public Commute Dataset: To better evaluate the number of commuters, we use two different datasets to evaluate the public commuters (*i.e.* MTA subway dataset, and NYC yellow taxis dataset). The combination of estimating commutes flow has been used and built a popular data visualization website[5]. The description of these datasets is shown below.

- **MTA Subway Dataset.** MTA subway dataset is published at MTA's official website[6]. Since the crime dataset is from 2014-2015, we downloaded the MTA subway dataset from 2014-2015. The total size of this dataset is about (900MB), including 11 fields. The attributes that we care about are station id, timestamp, number of entries, number of exits, station coordinate.
- **NYC Yellow Taxis Dataset.** Yellow taxis dataset covered Manhattan and this dataset contains detailed timestamp and number of passenger. We retrieve this dataset from NYC open data[7]. This dataset contains 19 fields. The total size of this dataset is about 25GB. The attributes that we care about are the timestamp, the number of passengers, pickup coordinate, drop off coordinate.

Since the format of MTA subway dataset and yellow taxis dataset mentioned above are similar, we can safely conduct the spatial (*i.e.* coordinate and precinct number) and temporal (*i.e.* weekday and hour) information in the same way. The datasets will combine and process as two datasets: (1) processed data#1 contains weekday, hour, precinct number, and the number of commuters; (2) processed data#2 contains weekday, hour, coordinate, and the number of commuters.

3) Twitter Dataset: The twitter dataset that we had used in this project is extremely large. The total size of the data is 313.7GB. This dataset[8] contains useful information like: timestamp, context, user, location. To process this dataset, we read the data chunk by chunk because of its large size. Then we clean up the data and use the NLP-based algorithm to calculate sentiment using the context provided. Eventually, we aggregated the data by weekday and hour which eventually conduct the processed data, including weekday, hour, number of positive, number of neutral, number of negative information. It worth mentioning that we didn't find the twitter dataset cover the whole year. We found the data covered Feb, Mar, Apr, May, Oct, Nov, and Dec in 2014.

IV. ALGORITHM

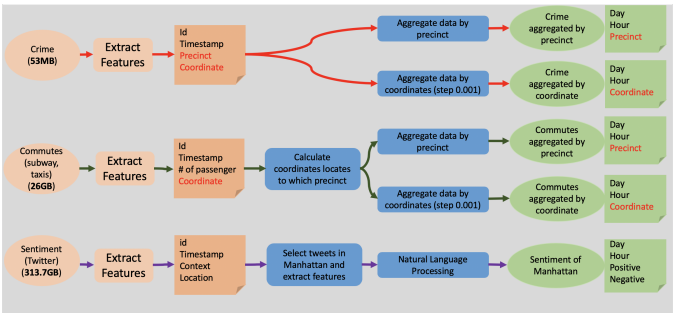

Fig. 3. Data pre-preocessing and processing workflow: (1) the shapes in orange demonstrate how we extract useful data from collected dataset; (2) the shapes in blue show the procedure of processing different dataset; (3) the shapes in green represent the final results after processing the data.

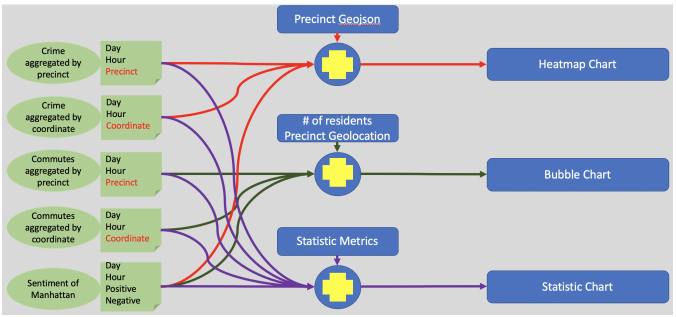

Fig. 4. Data visualization workflow: (1) we combined three dataset and NYC precinct geojson to build heatmap chart which aims to explore the spatial relationship; (2) we combined three dataset to build bubble chart which aims to discover the linear relationship; (3) we combine all final data to statistic chart whose goal is to summarize all statistic data to conduct the final result.

In this session, we will describe the detailed information regarding our algorithm to generate the final data which is further used for data visualization.

[4]https://www.kaggle.com/adamschroeder/crimes-new-york-city#Crime_Column_Description.csv

[5]http://manpopex.us/

[6]http://web.mta.info/developers/turnstile.html

[7]https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n

[8]https://archive.org/details/twitterstream

We will provide detailed information using a different algorithm to build generate the final data in Section IV-B. Lastly, in Section IV-C, we will discuss our idea to visualize the data which helps to explore the relationship of crimes, commuters, and sentiment in Manhattan.

It is worth mentioning that some datasets are really large (*i.e*. yellow taxis dataset, and twitter dataset), we use pandas to read the dataset chunk by chunk. In the rest of this session, we will read the dataset in fixed proper chunk size.

A. Data Pre-processing

In this session, we will describe the idea of how we pre-process the data, including extract useful data and clean up data.

1) Feature Extraction: We need to decide which data in the collected dataset are useful in this project. Since our goal is to explore the relationship among these datasets in Manhattan, we need geography location information, timestamp, and corresponding value. After preliminary analysis the data, we found all data have location information, timestamp, and specific value. We list that information below.

- **Crime Dataset.** It has coordinate, precinct number, the timestamp in the dataset, each column represents a criminal record.
- **NYC Yellow Taxis Dataset.** It records coordinate, timestamp, and the total number of passenger in the dataset.
- **MTA Subway Dataset.** It stores coordinate, timestamp, total number of entry and the total number of exit in the dataset.
- **Twitter Dataset.** It contains context, timestamp and location information in the dataset.

Therefore, we only keep these three attributes in the dataset. The orange shapes in Figure 3 shows the input (*i.e*. original dataset) and output (*i.e*. data after feature extraction).

2) Data Cleaning: After selecting the useful data from original datasets, we clean up the data which help us to get around some errors, such as missing data and invalid data. Our methods are listed below.

- **Remove duplicate rows.** Some of the rows in the dataset may store exactly the same information. We find those line and remove duplicates.
- **Remove rows with missing data.** Since the unused data columns have been removed, we are safe to remove the rows with missing data in the new data.
- **Fixed formatting.** Some float data (*e.g*. latitude, and longitude) store as the string in the dataset. We correct the format which makes easier for us to further utilize the data.
- **Round data.** Some float data (*e.g*. latitude, and longitude) stores the data in the different number of a decimal. We later will try different stepsize (*e.g*. 0.001, 0.005, 0.01) aggregated the data by coordinate. Round the data in this step help to reduce the size of data.

B. Data Processing

In this section, we will illustrate the detailed information on processing the data which eventually generate the data for visualization and exploration. The shapes in blue of Figure 3 shows how we process each dataset and eventually generate the final data for visualization.

To eventually generate the final dataset for visualization, we generate spatial and temporal indices (*i.e*. location and time) for data aggregation. We will give specific detail below.

- **Spatial.** For spatial, we use coordinate, precinct number, and Manhattan to aggregate the data because we are considering the flow/density of the data. Note that, we only consider twitter dataset of in Manhattan because of the limitation of the dataset. As a result, we consider the location using these three filters: (1) **Coordinate.** Since our goal is to explore the relationship of crime, commuters, and sentiment in Manhattan, we consider coordinate information which provides spatial information for exploration. (2) **Precinct number.** Precinct represents the district of a city or town as defined for police purposes and has a high correlation of crime [25]. Thus, besides exploring the data in coordinate, we also explore the data by precinct number. (3) **Manhattan.** The crime and commuter datasets we collected are focused on New York City, while the twitter dataset covers the whole United States. Therefore, we need to filter the data in Manhattan only.
- **Temporal.** As for temporal, we could further get the weekday, hour from the given timestamp, such that we can explore the temporal information hour by hour in an average week.

We will discuss the procedure of each dataset regarding the indices discussed above in the following three sub-sections.

	A	B	C	D	E
1	weekday	hour	crimes	lat	long
2	0	0	1	40.4998	-74.2449
3	0	0	1	40.5008	-74.2419
4	0	0	1	40.5058	-74.2499
5	0	0	1	40.5118	-74.2489
6	0	0	1	40.5138	-74.2449
7	0	0	1	40.5148	-74.2439
8	0	0	1	40.5228	-74.1849
9	0	0	1	40.5248	-74.1749
10	0	0	1	40.5258	-74.2019
11	0	0	1	40.5278	-74.2169
12	0	0	1	40.5318	-74.2229
13	0	0	1	40.5328	-74.1969
14	0	0	1	40.5338	-74.1529
15	0	0	1	40.5358	-74.2049
16	0	0	1	40.5358	-74.1829
17	0	0	1	40.5358	-74.1709
18	0	0	1	40.5388	-74.2399
19	0	0	1	40.5388	-74.1599
20	0	0	1	40.5388	-74.1459
21	0	0	1	40.5408	-74.1479
22	0	0	1	40.5418	-74.2029
23	0	0	3	40.5418	-74.1959
24	0	0	1	40.5418	-74.1499
25	0	0	1	40.5428	-74.1639

Fig. 5. Processed crime data: the total number of row is 764,523.

	A	B	C	D
1	weekday	hour	precinct	crimes
2	0	0	1	60
3	0	0	5	124
4	0	0	6	152
5	0	0	7	147
6	0	0	9	194
7	0	0	10	132
8	0	0	13	168
9	0	0	14	159
10	0	0	17	136
11	0	0	18	165
12	0	0	19	165
13	0	0	20	140
14	0	0	22	54
15	0	0	23	143
16	0	0	24	166
17	0	0	25	139
18	0	0	26	83
19	0	0	28	90
20	0	0	30	99
21	0	0	32	138
22	0	0	33	115
23	0	0	34	172
24	0	0	40	134
25	0	0	41	79

Fig. 6. Processed crimes data: the total number of row is 12,769.

	A	B	C	D	E
1	weekday	hour	commuters	lat	long
2	0	0	5	40.5478	-74.2199
3	0	0	1	40.5588	-73.9259
4	0	0	1	40.5608	-73.9109
5	0	0	1	40.5748	-73.9709
6	0	0	1	40.5758	-73.9809
7	0	0	2	40.5758	-73.9629
8	0	0	1	40.5788	-73.9679
9	0	0	1	40.5798	-73.9719
10	0	0	1	40.5808	-73.9819
11	0	0	1	40.5818	-73.9599
12	0	0	1	40.5828	-73.9599
13	0	0	1	40.5828	-73.9569
14	0	0	1	40.5828	-73.9539
15	0	0	6	40.5838	-73.9689
16	0	0	1	40.5838	-73.9659
17	0	0	1	40.5838	-73.9629
18	0	0	1	40.5838	-73.9529
19	0	0	5	40.5838	-73.9489
20	0	0	1	40.5838	-73.9359
21	0	0	1	40.5848	-73.9339
22	0	0	2	40.5858	-73.9539
23	0	0	2	40.5858	-73.9509
24	0	0	1	40.5858	-73.9319
25	0	0	1	40.5858	-73.8169

Fig. 7. Processed commuters data: the total number of row is 1,497,764.

1) Crime Dataset: As for crime dataset, we will list the detailed information of processing spatial and temporal information below.

- **Spatial.** Since crime dataset we collected already record detail location, including coordinate and precinct number. Therefore, as Figure 3 shows, we simply aggregate the data by coordinate and precinct number. (1) As for coordinate, we use a fixed step size (*i.e.* 0.001) to aggregate the data, because we care more about the density of each area. (2) As for the precinct number, we simply aggregate the data based on the same precinct number.
- **Temporal.** Since crime dataset records UTC timestamp, it is easy to convert UTC time to weekday and hour. Therefore, we simply use the datatime library[9] in Python to get the result.

Eventually, we generated two processed crime dataset: (1) one contains weekday, hour, coordinate, and the number of crimes information (Figure 5); (2) another contains weekday, hour, precinct number, and the number of crimes information (Figure 6).

2) Public Commute Dataset: Since commute datasets (*i.e.* MTA subway dataset, and NYC yellow taxis dataset) contain the data in exactly the same ways, the data processing are highly similar. We list the details of processing spatial and temporal information below.

- **Spatial.** Since commute dataset we collected only record coordinate information. Therefore, as Figure 3 shows, we further need to calculate the precinct number according to its coordinate. For this purpose, We follow the algorithm to calculate if a given coordinate is inside a polygon[10]. As each polygon represents the contour of the corresponding precinct number, we, therefore, calculate the precinct number of that record. Then we aggregated the spatial information. (1) As for coordinate, we use a fixed step size (*i.e.* 0.001) to aggregate the data, because we care more about the density of each area. (2) As for the precinct number, we simply aggregate the data based on the same precinct number.
- **Temporal.** Similar to crime data, since commuter dataset records UTC timestamp, it is easy to convert UTC time to weekday and hour using Python.

[9]https://docs.python.org/3/library/datetime.html

[10]http://alienryderflex.com/polygon/

	A	B	C	D
1	weekday	hour	precinct	commuters
2	0	0	1	1000
3	0	0	5	24153
4	0	0	6	65989
5	0	0	7	10810
6	0	0	9	40202
7	0	0	10	24735
8	0	0	13	42836
9	0	0	14	78088
10	0	0	17	48117
11	0	0	18	61975
12	0	0	19	24861
13	0	0	20	16547
14	0	0	22	3353
15	0	0	23	5754
16	0	0	24	10576
17	0	0	25	2043
18	0	0	26	3194
19	0	0	28	2652
20	0	0	30	1542
21	0	0	32	1111
22	0	0	33	1283
23	0	0	34	829
24	0	0	40	365
25	0	0	41	31

Fig. 8. Processed commuters data: the total number of row is 12,769.

Eventually, we generate two processed commute data: (1) one contains weekday, hour, coordinate, and the number of commuters information (Figure 7); (2) another contains weekday, hour, precinct number, and the number of commuters information (Figure 8).

3) Twitter Dataset: Twitter dataset records its data in different ways as the data in crime and commuter dataset, including the following challenges:

	A	B	C	D	E
1	weekday	hour	# of negative	# of neutral	# of positive
2	0	0	815	3604	1067
3	0	1	431	2297	767
4	0	2	462	2363	725
5	0	3	680	3140	872
6	0	4	907	3681	1026
7	0	5	919	4074	1095
8	0	6	922	4144	1141
9	0	7	930	4194	1001
10	0	8	893	3731	875
11	0	9	874	3730	850
12	0	10	718	2762	682
13	0	11	416	2016	388
14	0	12	420	1624	330
15	0	13	386	1509	319
16	0	14	302	1267	259
17	0	15	362	1404	295
18	0	16	388	1404	307
19	0	17	368	1391	314
20	0	18	334	1315	278
21	0	19	397	1447	487
22	0	20	605	2698	700
23	0	21	812	4341	1688
24	0	22	988	5598	2250
25	0	23	1041	5981	2475

Fig. 9. Processed twitter data: the total number of row is 169.

1) It records the data as JSON files.
2) It contains the tweets of the whole United States.
3) It is a large dataset whose total size is 313.7 GB.
4) It requires the NLP-based algorithm to calculate the sentiment.

We will list the detailed information below regarding the problem above.

- We follow the ETL pipeline to iterative find the useful data (*i.e.* timestamp, location, and context). Then output the data as CSV file. This solves the challenge 1) and 2).
- As challenge 3) described, the dataset is extremely large, we read the data chunk by chunk. Thus, this solves the challenges 3).
- Finally, we solve challenge 4) and generate **temporal**, **spatial** and **sentiment** information using the following steps. (1) We follow similar steps in crime and commuter dataset to convert UTC time to weekday and hour information. (2) Since we filter the location information by Manhattan, all the data left are posted in Manhattan. (3) Before process the context, we further process the context: remove the emoji context and replace the abbreviation with the corresponding full word. We use TextBlob[11] to calculate the sentiment of that tweet.

Finally, we generate the Twitter sentiment data as Figure 9. It includes temporal (column weekday and hour), spatial (since all the data are in Manhattan, we didn't specify this in the CSV file), and the number of tweets in different sentiment (column of positive, of neural, and of negative).

C. Data Visualization

Figure 4 shows the workflow of data visualization. It shows how we use different main dataset processed with the ancillary dataset to build the three charts (*i.e.* heat map chart, bubble chart, and statistic chart). It worth noting again that we eventually build up a website to support the public access.

[11]https://github.com/sloria/textblob

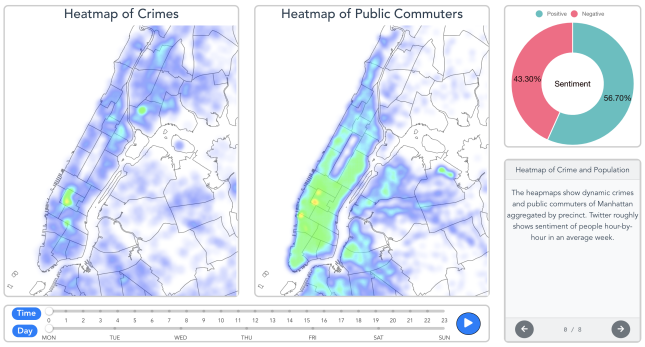

Fig. 10. Heatmap chart, including four parts: (1) heatmap of crimes and commuters in Manhattan (top left); (2) sentiment donut chart (top right); (3) day and hour slider bars (bottom left); (4) story text which highlight specific time (bottom right).

In the rest of this section, we will illustrate the goal and detailed design of each chart below. We will later discuss the result of each chart in Section VI.

1) Heatmap Chart: The **goal** of heat map chart is of showing spatial information regarding crime and commuters distribution in Manhattan. In addition, the sentiment of Manhattan people. Also, the whole chart should be interactive. Therefore, we design the page as Figure 10 shown. The **design** of this chart is listed below.

- We draw a Manhattan geography map and heatmap for crime and commuters. The brighter color of the heat map is the highest density of that area. The map is drawn as an SVG, using D3.js to add paths in Manhattan geography JSON file which is downloaded from NYC opendata[12].
- We add a donut chart at up right corner to demonstrate how sentiment changes in different time. We utilized and modified the donut chart in chart.js[13] to support the percentage of positive and negative sentiment.
- We later add two slider bars for the user to visualize the data in different time. Since we are using Vue.js, we found a useful slider bar in Github[14].
- Eventually, since a good visualization should let user discover it themselves [26]. We added a story text at the bottom right corner in which we highlight some specific moment to help reach a similar conclusion.

2) Bubble Chart: The **goal** of bubble chart is of showing the linear correlation of crime and commuters in Manhattan. The other features are similar to the heat map chart that we had the discussion at Section IV-C1. Figure 11 demonstrates our design of this chart. The detailed **design** is listed below.

- We draw a bubble chart using the data aggregated by precinct number. We use the number of the commuter at X-axis and number of crimes as Y-axis. Each bubble

[12]https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz
[13]https://www.chartjs.org/docs/latest/charts/doughnut.html
[14]https://github.com/NightCatSama/vue-slider-component

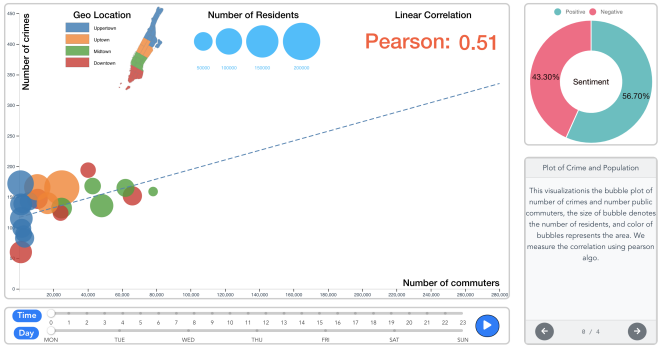

Fig. 11. Bubble chart, including four parts: (1) bubble chart of crimes and commuters in Manhattan (top left); (2) sentiment donut chart (top right); (3) day and hour slider bars (bottom left); (4) story text which highlight specific time (bottom right).

represents the corresponding data of a precinct. The color of the bubble denotes the geographic location of that precinct. The size of the bubble represents the number of residents in that precinct. We measure the linear correlation using Pearson Correlation which is shown at the up right corner of this chart. It worth mentioning that we split Manhattan as the upper town, uptown, midtown, and downtown, as it is described in Wiki[15].

- We add a donut chart at up right corner to demonstrate how sentiment changes in different time. This donut chart performs exactly the same as the donut chart in heat map chart.
- We add two slider bars for the user to visualize the data in different time. The slider bars perform exactly the same as the slider bar in heat map chart.
- Holding the similar idea as heatmap chart, we add the story text at the bottom right corner to helper user to explore the data.

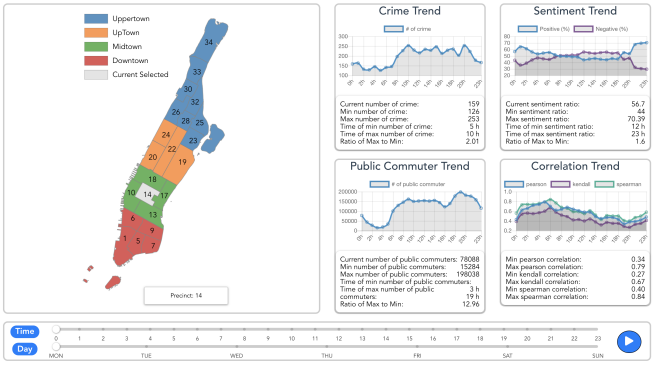

Fig. 12. Statistic chart, including four parts: (1) Manhattan precinct map (top left); (2) daily crime trend and commuter trend of each precinct (middle); (3) daily sentiment trend and correlation of Manhattan (right); (4) day and hour slider bars (bottom).

3) Statistic Chart: The **goal** of statistic chart is of summarizing all the data we had and eventually reach the conclusion of the relation among the data. 12 shows our idea of designing. The detailed **design** is listed below.

- We draw a precinct map on the left which using the exactly the same geojson file as we had used in heat map chart. This map is interactive. The user can choose different precinct to explore the crime and commuter trend via the line chart in the middle.
- We draw two line chart in the middle of this page. It shows the trend of crime and commuters each day.
- We draw another two line chart on the right-hand side of this page to show the trend of the sentiment of correlation each day.
- We add two slider bars for users to visualize the data in different time.

D. Summary of Algorithms and Tools

1) Algorithms: The algorithms that we had used are listed below.

- **Natural Language Processing.** We use NLP-based algorithm to process twitter dataset to get sentiment value. We use *TextBlob* which is a Python package of using *NLTK* and pattern to process context.
- **Calculate Point inside/outside of Polygon.** We use this algorithm to convert coordinate to precinct number.
- **Calculate Geo Center of Polygon.** We use this algorithm to put the precinct number of the map which provides better visualization effect.
- **Linear Regression.** We use linear regression to fit the model in bubble chart to explore the linear relationship of the number of crime and number of commuters.
- **Correlation Algorithms.** We use three different correlation algorithm (*i.e.* Pearson, Kendall, and Spearman) in statistic to better evaluate the correlation of data.

2) Packages and Tools: The packages and tools that we had utilized are summarized below.

- **Google Cloud Platform.** In order to process the really large amount of dataset, we use this to store and process data.
- **Pandas.** We use this package to process a large amount of data chunk by chuck at a relatively fast speed.
- **Numpy.** We use Numpy to boost the speed of matrix computation.
- **Shapely.** Shapely is a Python package for manipulation and analysis of planar geometric objects. We use it to calculate coordinate point inside/outside precinct polygon.
- **TextBlob** TextBlob is a Python package of using *NLTK* and pattern to process context. We use it to process Twitter context and calculate sentiment polarity.
- **Vue.js.** We use Vue.js to design the website because it supports the definition of the component which could be reused in different pages.
- **D3.js.** We use powerful functions of D3.js to manipulate DOM in HTML based on data.

[15]https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods

V. Software Package Description

In this section, we will show the usage of our data visualization website. The source code could be found in Github[16]

A. Data Preparation

In this section, we will summarize the data processing for those who want to reproduce from the very beginning. It is worth noting that we provide the processed dataset in our open source repository which can be directly used.

- **Collect data.** Collect each dataset from the corresponding website. The website and data description could be found in Section III-B.
- **Data Processing.** Process the dataset with corresponding algorithms. The detailed data processing steps could be found in Section IV-A and Section IV-B. It is worth mentioning that we utilize spatial (*i.e.* coordinate and precinct number) and temporal (*i.e.* timestamp) to process and generate new datasets that we could use for visualization.

B. Environment

- **Data processing.** To run our data processing scripts, you need to install **Python**. We use the packages *Pandas*, *Numpy*, *sklearn*, *matplotlib*, *TextBlob*, *shapely*, and *json*. We highly recommend to use *Google Cloud Platform* to process the data since the size of the dataset is large.
- **Data visualization.** To visualize the data, we had built a website. We had install *npm* to manager the packages. We use Javascript libraries to build the website, such as: *vue.js*, *D3.js* and *vue-slider-component*.

C. Usage

After processing the data and settling down all the environment, you are able to run all the codes and visualize the data in the website. We use **npm** to manager all the packages:

1) **Install npm packages.** Run the following command in the terminal.

```
cd <repository folder>
npm install
```

This command will install all the package that we had defined in *package.json*.

2) **Run server.** There are majorly two scenarios of running the server: (1) run and debug locally; (2) run and deploy in the cloud (*e.g. Google Cloud Platform*). Firstly, if you want to run it locally, you could run the following command.

```
cd <repository folder>
npm run dev
```

If you want to run it in *Google Cloud Platform* and make the website public. You could run as followed.

```
cd <repository folder>
HOST='0.0.0.0' PORT=8080 npm start
```

[16]https://github.com/zhichengMLE/relationship-among-crimes-commuters-and-sentiment-in-manhattan

VI. Experiment Results

A. Heatmap Chart

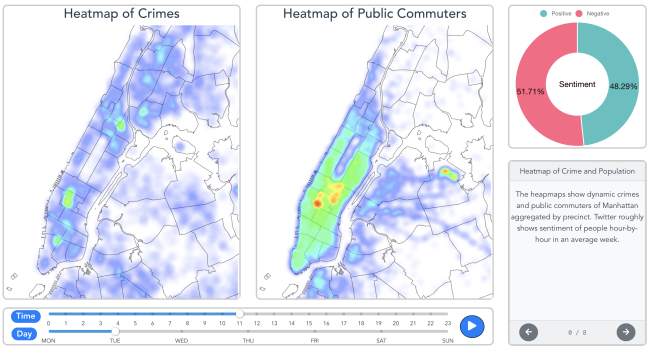

Fig. 13. Heatmap chart result on 11 am Wendesday.

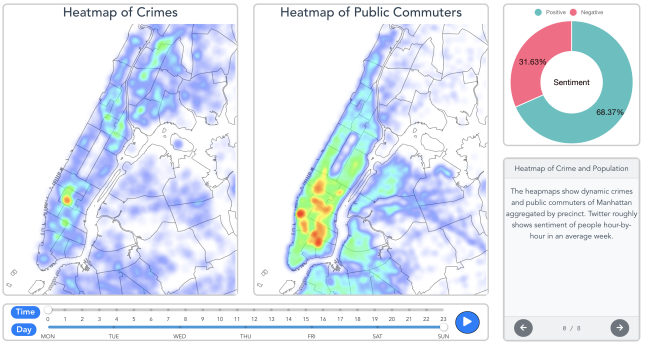

Fig. 14. Heatmap chart result on 0 am Sunday.

Figure 13 shows heatmap on 11 am Wednesday. We found that during workday hours, crimes diverse around a city with more density in the midtown area. In the meantime, commute population is quite low, because most the people are in the office. As a result, there is less chance for crimes happen.

Figure 14 demonstrates the heatmap on Sunday midnight. We notice that during a weekend night, crimes are hot in downtown, midtown and uptown area. As people commute mainly around the time square and hell kitchen neighborhood. Since at night, people leave the office and have dinner in the neighborhood where have a lot of restaurants and bars. When people try to relax, their attention is less focusing. So the crimes increase sharply. The place people are hanging out is also the place crimes happen a lot.

From the heat map chart, we notice that there are more commuters and crime in downtown and midtown while crimes are more diverse around the city. During the peak time in the weekdays, the crimes concentrate on the middle town and uptown area. As for the time out of the rush hour, the

crime rate and commute population are comparatively low and diverse around the whole city.

B. Bubble Chart

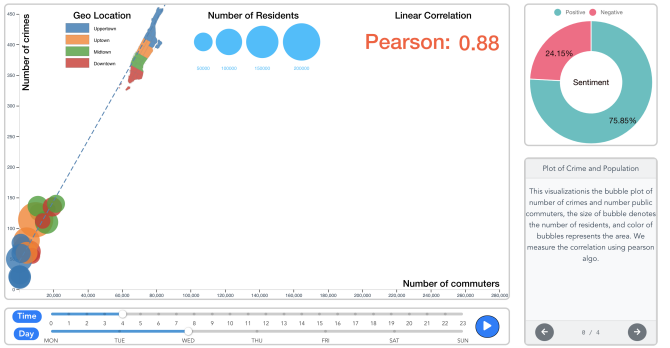

Fig. 15. Bubble chart result on 4 am Wednesday.

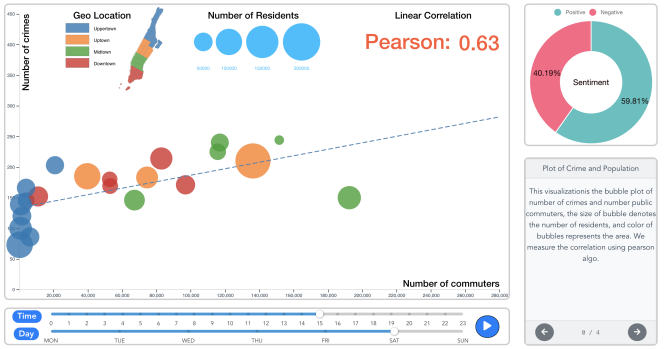

Fig. 16. Bubble chart result on 3 pm Saturday.

For the bubble chart, different bubble colors represent the different region of Manhattan (*i.e.* downtown, midtown, uptown, and upper town) and the size of bubbles indicate the population size of corresponding precinct.

Figure 15 shows the bubble chart result on 4am Wednesday. We notice that during the working hours among the weekdays, the Pearson correlation has a large value, indicates that the crime happens at the place with high compute density. The criminals are likely to select the neighborhood with high commute rate. When people are more crowded, the criminals have more chance to implement their actions. So the place with the high commute is also the place with high crimes.

Figure 16 demonstrates the bubble chart result on 3pm Saturday. Although the value of Pearson correlation is low than workdays, it is still above 0.5. So there is a tight relation between the crimes and compute density. The crimes are more diverse during this time. They won't select a specific place because there are lots of clusters of people all around. It is the best time for them. When people are less careful, they have more chance to arrive their destinations and goals.

To summary, downtown, midtown, and uptown have a relatively high frequency of criminal and high traffic density during work hours. It reduces gradually at midnight. From the Pearson correlation calculated, we can see that the number of crimes and commuters has strong linear positive relationship.

C. Statistic Chart

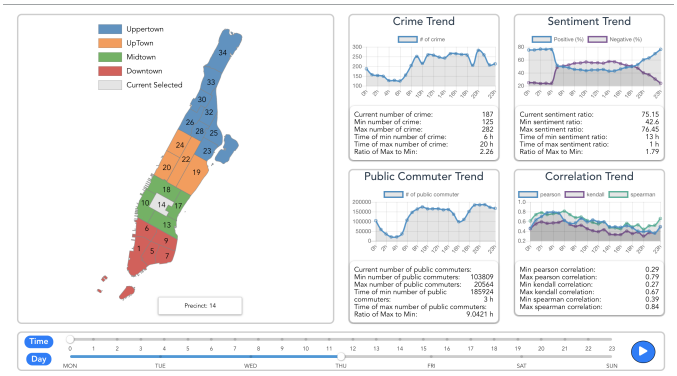

Fig. 17. Statistic chart result on Wednesday.

For the statistic chart, we can see how people move, and the relation in the whole week. Because the population is more focused in the midtown area. For example, on Wednesday, there is a strong relationship between the commuter, crime. People normally more negative for the work hours and positive for the night. In the weekend, people would be more positive all day.

During the weekdays, the city has relatively low commuter. Especially in the office hours from morning to evening, the population is quite low. People are a more likely site in the office room and not move around. During the office hours, people are always in an upset mood. We can suggest that people are always feeling stressed during work. They have to spend all their time and their energy to solve the problems they are facing in the office.

Figure 17 shows the statistic data on Wendesday. Since people are more concentrating on the midtown area, so we pick the precinct 14, midtown to illustrate. We found in our sentiment data comparison, people feel more negative during working hours, where the negative percentage is quite high in the value of nearly 70 percent and feels positive during off-hours, like night and weekends. It is also a time for crimes happen. Criminals have more opportunities during this time, So the number of commuters increase as well as number of crimes.For the relationship between crimes and commuters data, we use three different algorithm to test our hypothesis, which are pearson, kendall, spearman. All three methods show similar results which backs up the conclusion that there is a tight positive relationship between crimes and commutes.

VII. Conclusion

From the heat map charts, we notice that there are more commuters and crimes in downtown and midtown, while less in the uptown and upper town. From the Pearson correlation calculated, we can see that the number of crimes and commuters are basically positive correlated (above 0.5). From the

sentiment analysis, we found that for the weekdays, people are more negative during work hours and they tend to feel more positive after work. But for the weekend, people would be more positive all the day which may result in safety awareness decreasing and lead number of crimes increase. There are also some special events during in people's life. The Wednesday is the saddest day during the whole week, especially during the daytime. There is very limited commute while crimes drop a little. People have the most negative mood. The weekend, including the Friday night, has the highest positive feelings. During this time, crimes increase sharply while the city has a high commuter population. We can make the conclusion that, the crimes have a tight relationship between the city commute population and people sentiment polarity. When the commute population increase, the number of crimes increase. When the sentiment is more positive, the more crimes happen.

ACKNOWLEDGMENT

The preferred spelling of the word is to our professor, CHING-YUNG LIN for the wonderful lecture. We appreciate the help from teaching assistants. Thanks to the Kaggle, NYC Open data, MTA, and Archive for the dataset.

APPENDIX

A. Individual Contribution

- Zhicheng Ding contributes on Crime dataset and Twitter dataset, including extremely large dataset processing and generates crime info and sentiment polarity from the dataset. He also works for data visualization and part of report.
- Yunan Lu has been working on processing the MTA subway dataset and NYC yellow taxis dataset. She also makes the PPT and part of report.
- Lin bai helps collect the dataset and explore the usage of TextBlob for Sentiment Analysis. She also makes the video and part of report.

REFERENCES

[1] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto, "Big data, bigger dilemmas: A critical review," 2015.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1007/s11036-013-0489-0

[3] G. Weikum, "Foundations of statistical natural language processing," *ACM SIGMOD Record*, 2002.

[4] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, 2013.

[5] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -, 1999.

[6] O. Nasraoui, "Web data mining," *ACM SIGKDD Explorations Newsletter*, 2008.

[7] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Systems*, 2014.

[8] M. Naaman, "Social multimedia: Highlighting opportunities for search and mining of multimedia data in social media applications," *Multimedia Tools and Applications*, 2012.

[9] M. Steketee, A. Miyaoka, and M. Spiegelman, "Social Network Analysis," in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 2015.

[10] L. Branz and P. Brockmann, "Sentiment Analysis of Twitter Data," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems - DEBS '18*, 2018.

[11] A. Bifet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[12] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: Using location data from cell phones for urban analysis," 2006.

[13] A. Ghose and S. P. Han, "User Content Generation and Usage Behavior on the Mobile Internet: An Empirical Analysis," 2010.

[14] J. Boase and R. Ling, "Measuring Mobile Phone Use: Self-Report Versus Log Data," *Journal of Computer-Mediated Communication*, 2013.

[15] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: A general framework and some examples," *Computer*, 2004.

[16] C. H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2011.

[17] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *Journal of Transport Geography*, vol. 43, pp. 78 – 90, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966692315000253

[18] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, ser. TDMA '11. New York, NY, USA: ACM, 2011, pp. 23–30. [Online]. Available: http://doi.acm.org/10.1145/2030080.2030086

[19] Y. Zhu, K. Ozbay, K. Xie, and H. Yang, "Using Big Data to Study Resilience of Taxi and Subway Trips for Hurricanes Sandy and Irene," *Transportation Research Record: Journal of the Transportation Research Board*, 2016.

[20] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, 2013.

[21] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, Dec 2012, pp. 182–188.

[22] W. Janyce, "Learning Subjective Adjectives from Corpora," *Proceedings of the National Conference on Artificial Intelligence*, 2000.

[23] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 355–363. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610125

[24] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 36–44. [Online]. Available: http://dl.acm.org/citation.cfm?id=1944566.1944571

[25] P. J. Levchak, "Do Precinct Characteristics Influence Stop-and-Frisk in New York City? A Multi-Level Analysis of Post-Stop Outcomes," *Justice Quarterly*, 2017.

[26] J. J. Thomas and K. A. Cook, "A visual analytics agenda," 2006.