ABSTRACT

Title of Thesis:	SPATIOTEMPORAL ANALYSIS OF PUBLIC SENTIMENT WITH TWITTER: A CASE STUDY IN NEW YORK CITY, USA
	Jinyi Cai, Master of Science in Geospatial Information Science, 2019
Thesis Directed By:	Dr. Eunjung Lim, Department of Geographical Sciences

Big data from social media can make a large profit and benefit daily life. However, current research and application focus on the descriptive analysis of the public sentiment lacking of the spatiotemporal pattern of sentiment polarity with quantitative analysis. In this capstone project, the spatial and temporal pattern of sentiment distribution of twitter data was analyzed to understand the spatial and temporal distribution of public happiness within New York City with a multivariate linear mixed-effect model. The twitter data was retrieved with Twitter Streaming API and cleaned with NLTK tool and regular regression. The Sentiment analysis was implemented with a deep learning model with LSTM neural network. The fixed effect of land use and time period to the sentiment score was investigated with a multivariate linear mixed-effects model. Finally, the tweets with sentiment scores were visualized as a Web GIS Application. The sentiment analysis result showed that half of the sentiment was positive (53.01%), and the other half were neutral (34.57%) and negative (12.42%). The result of the linear mixed-effect model indicated the significant effect of the time period, days of the week and land use type on the sentiment score. Among these fixed effects, the effect of land use categories was the largest. The result of the fixed effect revealed that the sentiment decreased by 0.084362 in the transportation area during late night on Friday than the sentiment in the recreation area before dawn on Saturday.

SPATIOTEMPORAL ANALYSIS OF PUBLIC SENTIMENT WITH TWITTER: A CASE STUDY IN NEW YORK CITY, USA

by

Jinyi Cai

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of Master of Science
in Geospatial Information
Sciences
2019

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

Nowadays, social media have become an important source of various real-time information. Twitter is one of the most popular social media platforms with 500 million tweets sent per day (2018). About 80% of twitter users tweet with mobile phones. Among them, users who tweet with geotags can provide abundant spatiotemporal real-time information like opinions and sentiment.

Big data from social media can make a large profit and benefit daily life. The sentiment analysis of social media has been applied in the marketing areas. Social media examiners like MomentFeed and Local Measure help companies analyze the local social data to create targeted marketing campaigns. They gain valuable insight from the current trend and user sentiment to understand their performance. The analytics provide competitive intelligence to help the companies measure their performance and return on investment (ROI) at each point of sale.

Literature Review

Many researches explore the application of Volunteered Geographic Information (VGI) from social media. Twitter has been a valuable source of VGI to provide information in fields such as stock markets, public health and opinion polling. The intrinsic knowledge extracted from tweets can help government and business intelligence.

Cao and MacNaughton investigated how land use and time influence the public sentiment by analyzing the sentiment of 880,937 tweets within Massachusetts from

November 31, 2012 to June 3, 2013 with the IBM Watson Alchemy API. Then they used the multivariate linear mixed effects model to evaluate the influence. They found that the sentiment scores were higher at public and commercial places, on weekends and at noon/evening (Cao et al., 2018). The powerful tool of IBM Watson Alchemy API provides Natural Language Understanding service for users to extract concepts, entities, key words, categories, opinions, sentiments, syntax rules and relationships from text. Research shows that the Alchemy APIs could be a suitable tool for analyzing any kind of text because of its high performance in both short and long texts. (Serrano-guerrero, Olivas, Romero, & Herrera-viedma, 2015).

Yang and Mu designed a procedure to detect users with depression from tweets they sent and analyzed their spatial patterns. They differentiated the tweets posted in New York City from September 5, 2013 to March 5, 2014 associated with depression from the non-depression tweets with NMF. Then they diagnosed the text with the DSM-IV criteria (Yang & Mu, 2015). This research explored a new diagnosis method in the field of public psychological health. However, the limitation was that the depression diagnoses from social media could not be highly reliable. It should be more confident by integrating the social media data and other data sources like professional clinical scale table.

Bollen and Mao investigated the correlation between the large-scale public sentiment from twitter and the value of the DJIA with 985,000 tweets collected from February 28, 2008 to December 19, 2008. They analyzed the sentiment with two mood tracking tools and cross-validate the result by comparing their capability of detecting the public response to the politic election and Thanksgiving Day. The result showed that the public mood could increase the accuracy of Dow Jones Industrial Average (DJIA) predictions (Bollen, Mao, & Zeng, 2011). The important factors that this research did not consider was that stock markets in the USA can be affected by the public sentiment around the world while the tweets they collected were written in English. Besides, they did not consider the location factors.

Song and Xia investigated the spatial and temporal distribution of user sentiment in a community with 3097 tweets collected within the Curtin University boundary. They analyzed the sentiment with Starlight Visual Information System and converted the non-spatial files into spatial and temporal features with Feature Manipulation Engine. The research found that the social science area had most positive tweets while the science, engineering and dormitory areas had most negative tweets. At the exam period, the negative tweets increased to the peak at the library, science, and engineering areas while dropping at the entertainment and sport and dormitory areas (Song & Xia, 2016).

Stojanovski and Chorbev detected and identified the social hotspot from Twitter and applied sentiment analysis. For the hotspot detection, they used hierarchical agglomerative clustering and Density-based spatial clustering of applications with noise (DBSCAN) to get the cluster pattern. Then they analyzed the text sentiment with the Convolutional Neural Networks (CNN) and depicted a sentiment heatmap. The result showed the spatial distribution of the positive, negative, and neutral sentiment. Besides, there is an overlap between positive and negative sentiment (Stojanovski & Chorbev, 2016). The limitation of this research was that they did not further their sentiment result to analyze the spatial and temporal pattern.

In the marketing areas, the analysis of social media has been applied widely. Social media examiners like brand 24 detect public sentiment for their company customers. They gain valuable insight from the current trend and user sentiment to understand their performance. The analytics provide competitive intelligence to help the companies measure their performance and return on investment (ROI) at each point of sale.

The researches above provide useful points in terms of data retrieval, study region and analysis method. For data retrieval, the twitter streaming API was used to collect the tweets data, which has been proven to be an official tool to acquire twitter data. For the analysis method, the spatial and temporal pattern of the results were analyzed. The multivariate linear mixed-effect model was used for analyzing the effect of land use and time period on the sentiment score.

Objectives

As the introduction above, intrinsic knowledge extracted from tweets can help government and business intelligence in a more dynamic and economic may compared with the traditional survey method. However, current research and application focus on the descriptive analysis of the public sentiment lacking of the spatiotemporal pattern of sentiment polarity with quantitative analysis.

In this capstone project, the spatial and temporal pattern of sentiment distribution of twitter data, the social media data, was analyzed to understand the spatial and temporal distribution of public happiness within New York City with a multivariate linear mixed-effect model.

Chapter 2: Data

Study Region

The study area of this project will be New York City (Figure 1). The twitter data sent in New York City was collected from September 20, 2019 to October 1, 2019. The bounding box of longitude and latitude was (−74, 40), (−73, 41). The language should be English for sentiment analysis.

New York City was chosen because this is one of the most active twitter places with high population density. Plunz et al. (2019) measured the public well-being by analyzing the sentiment of the tweets within and outside New York City parks (Plunz et al., 2019). Eldering streamed tweets in New York City to analyze public space identification (Eldering, 2017).

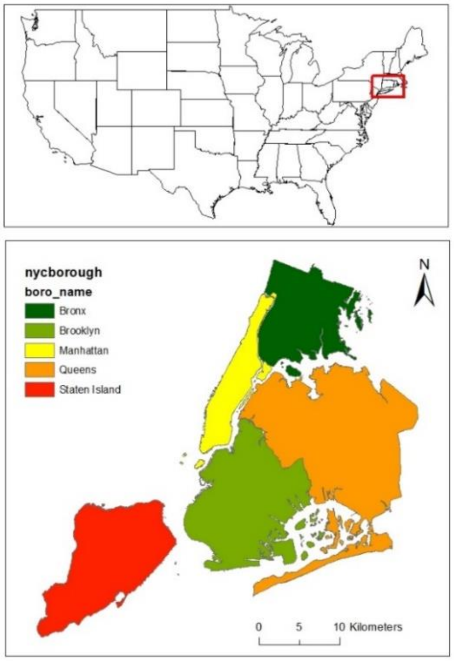


Figure 1 Study Region: New York City

Data

The detail descriptions of data sources, data quality and data usage are shown in Table 1.

Table 1 Data Description

Data	Date	Data Source	Quality	Format	Usage
Tweets	10/14/2019 - 10/21/2019	The original tweets data will be collected with Twitter Streaming API. https://developer.twitter.com/en/docs	Having noise and null value and needs to process with data cleaning	.json	The tweet content was used to calculate the sentiment score. The latitude and longitude will be used to add point vector with sentiment score into a map and further with spatial modeling
Sentiment140 dataset with 1.6 million tweets	2009	Stanford University Dataset downloaded from Kaggle dataset https://www.kaggle.com/kazanova/sentiment140	tweets have been annotated (0 = negative, 2 = neutral, 4 = positive)	.csv	It was used to build the sentiment classification model by splitting into training dataset (70%), validation dataset (10%) and test dataset (20%)
Land Use Zoning Districts for New York City	Published at 8/30/2019	NYC Department of City Planning, Technical Review Division https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-gis-zoning.page	Land use type Lack of transportation area	.shp	The zoning layer will be used to assign the land use type to tweet point layer with Spatial Join tool in ArcMap

NYC Planimetrics	Created at: 3/10/2016 Updated at: 2/1/2019	Department of Information Technology & Telecommunications https://data.cityofnewyork.us/Transportation/NYC-Planimetrics/wt4d-p43d	complete	.shp	The roadbed features of New York City were extracted from the planimetrics layer and union with the land use zoning layer

Data Collection

Twitter Data

72740 georeferenced tweets were collected with the Twitter Streaming API, sent in New York City with the bounding box of longitude and latitude of (−74, 40), (−73, 41). The result was the JSON object storing each tweet with total 87 attributes which details the tweet information including the create time, location with coordinate, tweet id, source, retweet, hashtag, etc., and the user profile including the user name, screen name, user id, user location with coordinate, etc. The percentage of data collected in a week from Monday to Sunday were 18.93%, 17.38%, 11.17%, 13.59%, 13.70%, 13.19%, 12.05% respectively. The percentage of tweets in different time period were 15.74% in the morning (6:00–11:00), 17.35% at noon (11:00–14:00), 22.72% in the afternoon (14:00–18:00), 21.92% in the evening (18:00–22:00), 11.62% at night (22:00–00:00), 4.56% at late night (00:00–3:00) and 6.09% before dawn (3:00–6:00).

160000 tweets annotated with sentiment label (0 = negative, 2 = neutral, 4 = positive) were downloaded from the Kaggle which were aimed to use as competition dataset to detect sentiment. The dataset contains 6 attributes: polarity, id, date, flag, user and text.

Land Use Zoning Districts for New York City

The land Use Zoning Districts layer consists of polygon features representing the zoning districts in New York City (figure 2). These features cover the entire city, extending to the city limits on land and out to the US Army Corps of Engineers' pierhead lines over water. The data were developed using Department of City Planning's Tax Block Base Map Files, Department of Information Technology & Telecommunications' NYC Map planimetric street centerlines and 2006 orthophotos as reference sources.

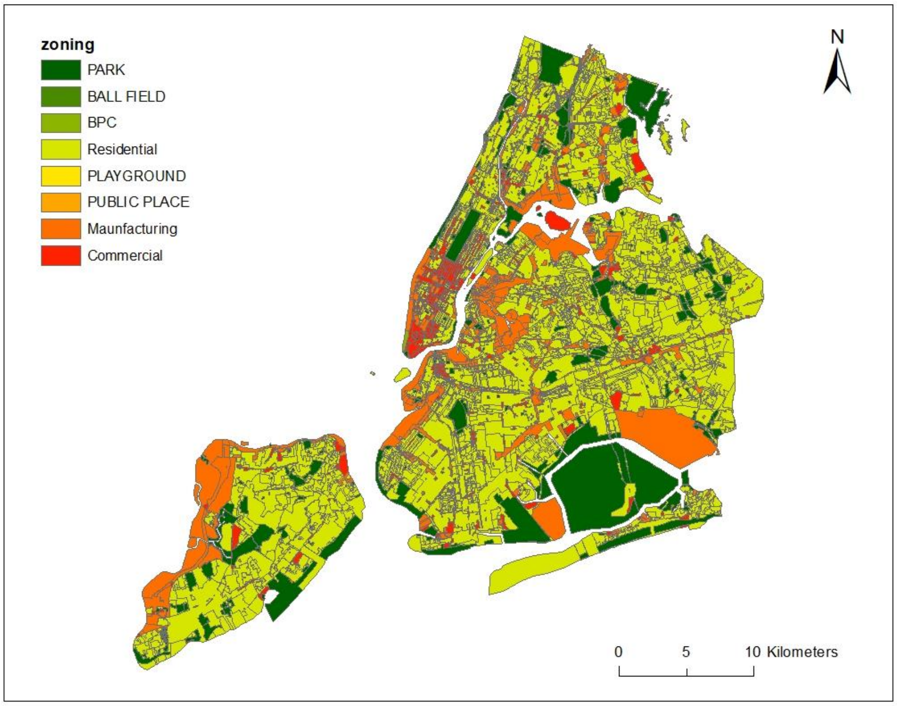


Figure 2 Land use zoning of New York City

NYC Planimetrics

Digital planimetrics were derived using the imagery products delivered with the 2014 New York City Statewide Flyover, which includes raw imagery collected to support the generation of 0.5 Ft Ground Sample Distance (GSD) natural color imagery. Roadbed represents the interior polygon of the pavement edge. The edges of these features were coincident with the linear feature class Pavement Edge.

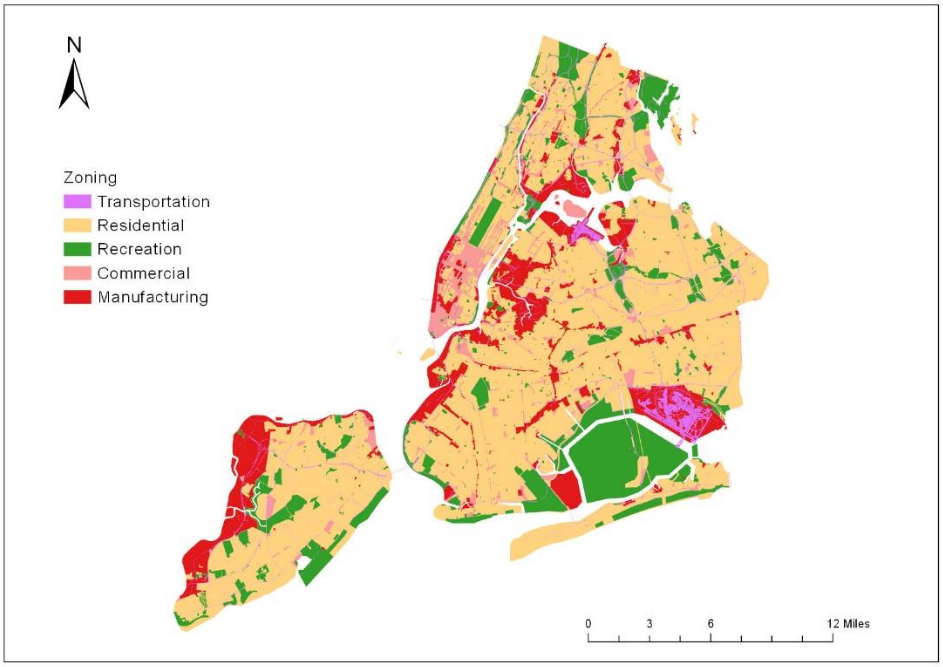


Figure 3 Land use categories of New York City

The eight land use types from the land use zoning data and roadbed features were grouped into five categories based on their intra-similarities (Table 2). Commercial areas included retails and service shops in neighborhoods and central business districts at the city center. Residential districts consisted of variety of residential buildings including the single-family homes on the outskirts of the city and the high-density apartment at the center of NYC. The recreation areas consisted of parks, ball fields, public places for pastime and playgrounds. The transportation areas consisted of roadbed, driveway, road shoulder and intersection extracted from the planimetrics layers. The manufacturing districts included traditional industrial areas and municipal facilities such as catering suppliers, lighting fabricators, warehouse, ferry and ship terminals.

Table 2 Land use Category of New York City

Category	Land Use Type	Area (km^2)
Commercial	Commercial, Special Battery Park City District (BPC)	41.1713
Residential	Residential	533.0148
Recreation	Park, Ball Field, Public Place, Playground	141.8695
Transportation	Roadbed, Driveway, Shoulder, Intersection	29.8058
Manufacturing	Manufacturing	114.2605

Chapter 3: Methodology

The programming language of this project was python. Most of the work was run with Jupyter Notebook, a web-based interactive programming environment. The flowchart of the methodology is shown in Figures 4, 7 and 8. The critical script code was appended at Appendix A.

Data Processing

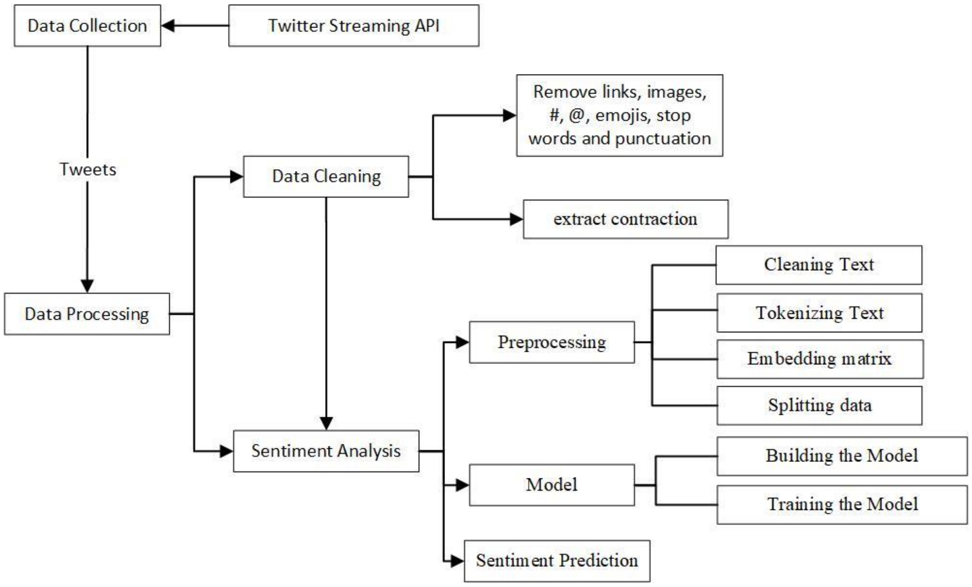


Figure 4 Flowchart of the data processing

Data Cleaning

The data cleaning process was necessary to get rid of the null values and noise in the source tweets data. The attributes of 'tweet_id', 'text', 'time', 'coordinates', 'user_id' and 'statuses_count' (indicating the number of Tweets issued by the user) were selected for the sentiment analysis. Then the data with null values were dropped. The noises, links, images, hashtags #, @ mentions, emojis and punctuation, were removed with regular expression. The stop words were removed with the stopwords() method from the Natural Language Tool Kit (NLTK) library. Phrases like "what's" were extracted to "what is' to extract contraction with regular expression. Words like “ran” were stemmed to “run” with stem() method from NLTK so they were all the same tense.

Sentiment Analysis

The Keras deep learning library tool was used to build the model to analyze the sentiment of the tweet text with the Long Short Term Memory (LSTM) network. The code was run on the Kaggle Kernel platform which provides free access to NVidia K80 GPUs to speed up the training of deep learning process. The model (Figure 5) took in an input of a tweet text and then outputted a float number within the range [0, 1.0]. The 0 and 1 meant extremely negative and extremely positive separately. The 0.5 value indicated neutral sentiment.

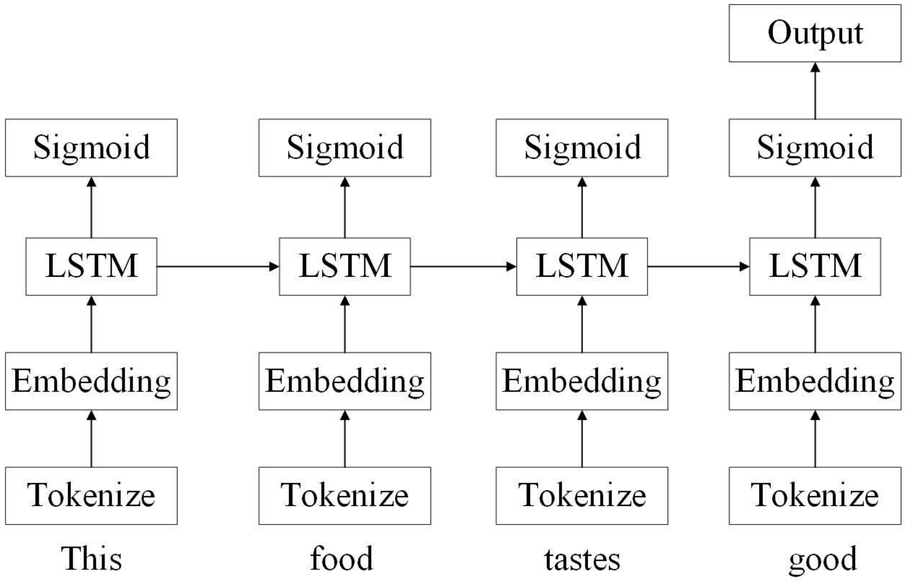


Figure 5 Architecture of the deep learning model

The deep learning model was built with a preprocessing layer of tokenizing and embedding layer. The tokenizing layer used a Tokenizer to assign indices to the words in the training data based on frequency.

In the embedding layer, the word embedding was built with the Word2vec algorithm. Each word was represented as a n x m embedding matrix where n is the number of words and m is the dimension of the embedding.

The LSTM layer (Figure 6), known as hidden layer, took in each input embedding vector. The LSTM is one of the most applicable neural networks for natural language processing (NLP). It is one of the variations of the Recurrent Neural Network (RNN). The RNNs are designed for detecting and generating time-varying patterns. They accept an input vector x and give an output vector y. This output vector's contents are influenced not only by the direct input, but also on the entire history of inputs in the past. The LSTM is built on the base of RNN which can not only process text with memory of each word but also solve the exploding and vanishing gradient problems so that it can process with long- and short-term memory.

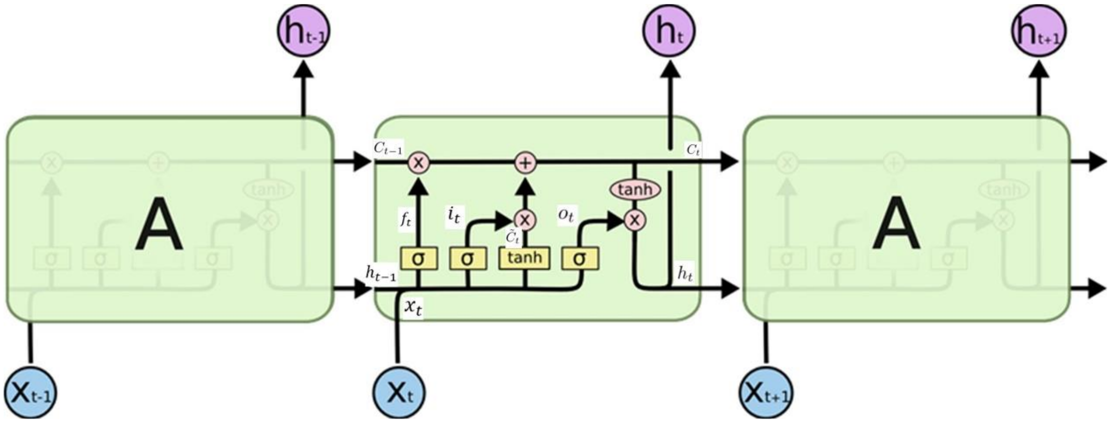


Figure 6 The repeating module in an LSTM network

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_f * [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C_t} = tanh(W_C. [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh{(C_t)}$$

Where:

h_t: recurrent vector to the next module

h_{t-1}: recurrent vector from the last module

x_t: input word vector

C_t: new cell state

C_{t-1}: old cell state

$\widetilde{C_t}$: the vector that could be added to the cell state

f_t: forget the gate layer to decide whether throw away the input from the cell state. Output of 1 means remain and 0 means throw away.

i_t: input gate layer to decide whether the new information will be stored in the cell state

o_t: output gate layer to decide what parts of the cell state is going to be the output

Then through the sigmoid layer the outputs from the LSTM layer are transformed between zero and one. The sigmoid output of the last word of the sentence was the final output of the model.

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

The model was built and fit with the Sequential() module of the keras library.

The accuracy evaluation was conducted with the validation data and test data. The validation data was used to evaluate the model fit to tune the hyperparameters. Hyperparameters are settings that can be tuned to control the behavior of a machine learning algorithm such as epoch, number of hidden units or the learning rate of a model. Since there is possible that the accuracy of the training model is better but on the validation becomes bad, which is overfitting, with the validation accuracy evaluation of each epoch, the optimal times of epoch can be recognized. After building and fitting the model, the test data was used to evaluate the accuracy of the final model.

Regression Analysis

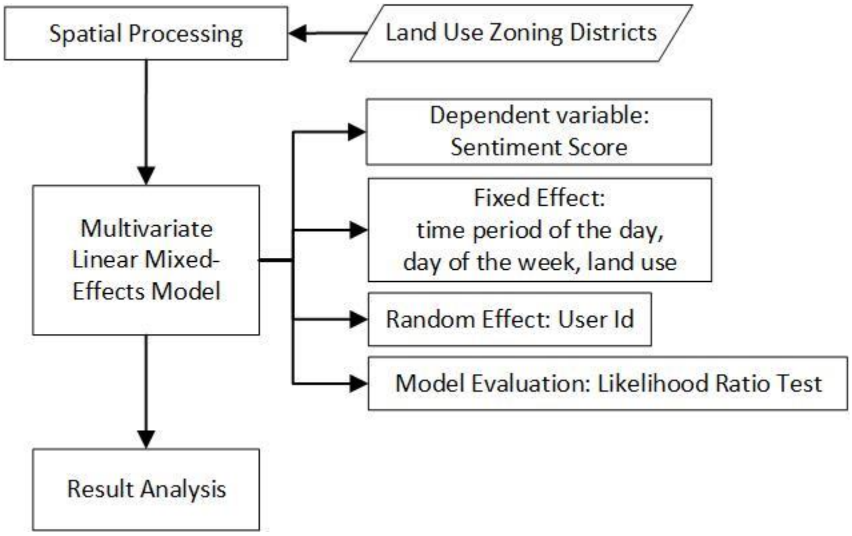

Spatial Processing
Land Use Zoning Districts
Dependent variable: Sentiment Score
Fixed Effect: time period of the day, day of the week, land use
Multivariate Linear Mixed-Effects Model
Random Effect: User Id
Model Evaluation: Likelihood Ratio Test
Result Analysis

Figure 7 Flowchart of the regression analysis

Spatial Processing

Use the Spatial Join tool in ArcMap to assign each tweet point layer with the land use zoning type. The create time of tweets were processed to convert from UTC time zone into America/New_York time zone and format the time date to acquire the data of period of the day and day of the week with attribute field calculator. The time periods of the day consisted of morning: 6:00–11:00, noon: 11:00–14:00, afternoon: 14:00–18:00, evening: 18:00–22:00, night: 22:00–00:00, late night: 00:00–3:00, before dawn: 3:00–6:00. The land use category consisted of the recreation area, commercial area, manufacturing area, transportation and residential area.

Regression Analysis

A multivariate linear mixed-effects model was built to analyze the effect of land use and time period on the sentiment score of different users. Since there were multiple measures with the same tweet users, which violate the independent assumption of linear model, the user id was used as the random effect to eliminate the influence. Based on the study of Xiaodong Cao (Cao et al., 2018), the analysis was performed with the “lme4” package in R with the regression model.

Since the time period were independent variables and divided into time steps, this is a dynamic and discrete model. It is stochastic because the vector of random errors ε represents the deviations from the predictions due to “random” factors that the model does not include. The model parameters derived from measured data with the linear regression algorithm so it is an empirical and linear model.

$$y_{i,j,k,l} = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_5 \times L_5 + \beta_6 \times D_1 + \cdots + \beta_{12} \times D_7 + \beta_{13} \times T_1 + \cdots + \beta_{18} \times T_7 + b_{0,i} \times \varepsilon_{i,j,k,l}$$

$y_{i,j,k,l}$: sentiment score for user i in land use category j during time period l on day k.

$b_{0,\, i}$: random effect of the intercept for user i to account for the clustering effect of each

L_j: land use category

D_k: categorical variable of the days of the week

T_l: categorical variable of the time periods of the day

β_1: fixed intercept

β_1–β_5: fixed effects of other land use categories compared to recreation area

β_6–β_{12}: fixed effects of other days of the week compared to Saturday

β_{13}–β_{18}: fixed effects of other time periods of the day compared to before dawn

$\varepsilon_{i,j,k,l}$: vector of random errors.

The Likelihood Ratio Test will be used for significant evaluation by comparing the model with the null model using anova function in R. The Analysis of Variance (ANOVA) is a statistical method used for the significant test of model with more than two variables.

Visualization

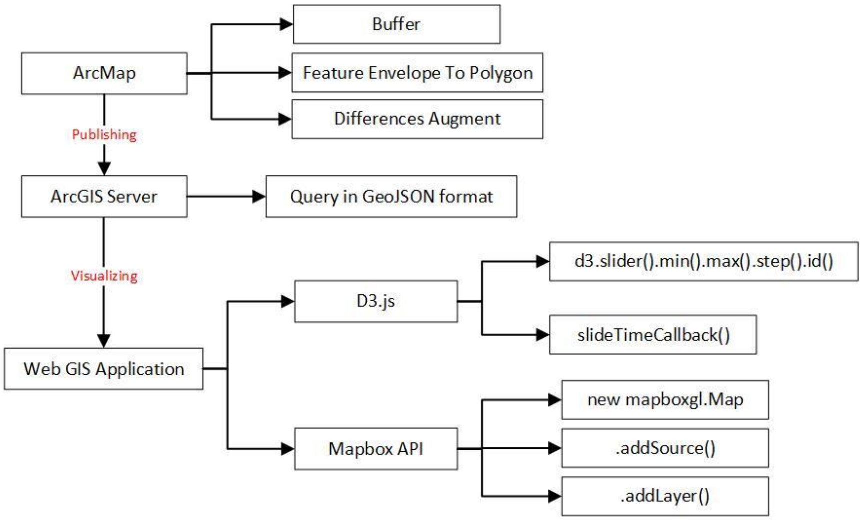


Figure 8 Flowchart of the visualization

Spatial Processing

The point layer was buffered with a distance of 100 meters and the feature envelop to polygon tool was used to get polygon features. The polygon layer was converted into GeoJSON format and published to the Mapbox Studio as a tileset file so that it could be added as a source layer with the Mapbox API.

Web GIS Application Building

The Mapbox GL JS is a JavaScript library that uses WebGL to render interactive maps from vector tiles and Mapbox styles. It is part of the Mapbox GL ecosystem, which includes Mapbox Mobile, a compatible renderer written in C++ with bindings for desktop and mobile platforms.

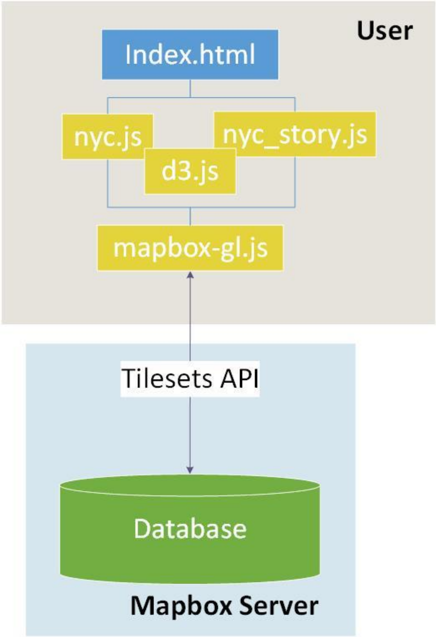


Figure 9 Architecture of the Web GIS Application

The new mapboxgl.Map was used to create a new map object. The map object exposed methods and properties to change the map programmatically, and fired events as users interact with it.

The on() method was used to initialize and call functions for the map. In this method, there were addLayer() and setFilter methods. The addlayer() added a Mapbox style layer to the map's style with the tileset source uploaded to the Mapbox studio. In the addlayer() method the fill-extrusion property could be used to create a 3D bar graph for better visualization.

The D3.js is a JavaScript library used to bind data to a Document Object Model (DOM), and then apply data-driven transformations to the document. The time range within the days of the week and hours of the day was implemented by drawing a slider with d3.slider() .

The slideTimeCallback() and slideendTimeCallback() function can return the time range to show the time change on the page and get the sentiment score. So did the change of the days of the week.

When the user changes the slider, the d3.select() method would be used to call the d3.slider() method

Once the slider was changed, the changeTime() function would be called to update the data shown on the map with the setFilter() method of the map object. The daytime variable was the combination of day and time value obtained from the slider of days of the week and time period of the day.

In the story mode, the slider changed with a different story by a updateStoryDaytime() function, which called the slideTimeCallback() function to update the data shown on the map with the setFilter() method at nyc.js. The sliders on the map were moved by the value() method of the sliderTime object.

The district box was check as the story page change with updateStoryDistricts() function with d3.select().property() method. The data on the map was updated with the setFilter() method .

Chapter 4: Results and Discussions

Model Result

The training data was used to train the model and the validation data was used to describe the evaluation of models when tuning hyperparameters and data preparation. After trained 15 epochs time, and the final validation accuracy of the model is 0.7840 (Figure 7).

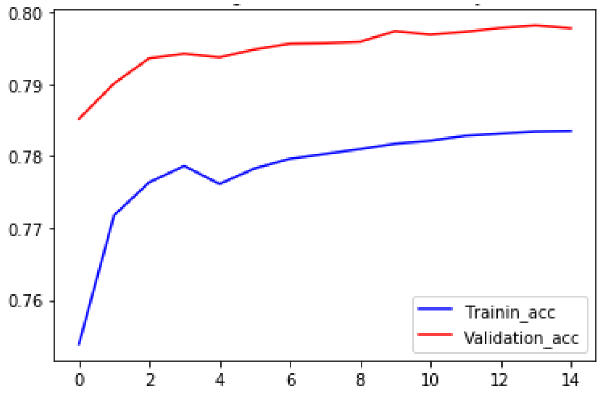


Figure 10 Result of the training and validation accuracy

The test data, 320000 tweets, was used to evaluate the accuracy of the model by comparing the true sentiment label and the predicted sentiment label and got the accuracy of 0.7972. Since the output of the model was a sentiment score range from 0 to 1, the sentiment label was set as 0: 0.0-0.4, 2: 0.4-0.6, 4:0.6-1.0, which 0, 2 and 4 means negative, neutral and positive sentiment. Figure 8 shown that half of the sentiment was positive (53.01%) and the other half were neutral (34.57%) and negative (12.42%).

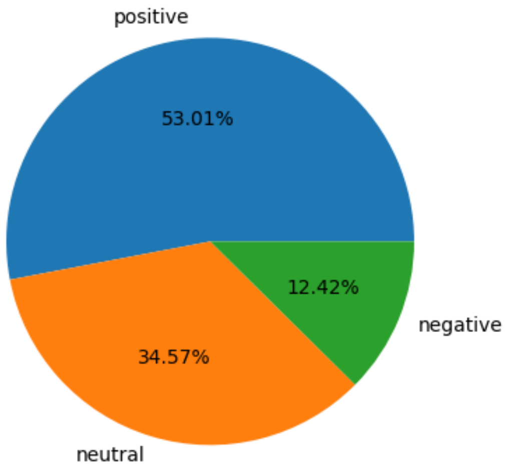


Figure 11 Sentiment Pie

Sentiment Distribution

43693 tweets were left in the NYC region after the data cleaning and spatial processing (Figure 9). Table 3 indicates that land use type with the highest density of the tweets was found in commercial areas with 425 tweets every square kilometer. The transportation areas had 122 tweets every square kilometer ranking the second highest density areas. The recreation areas and residential areas had relatively low density each with 23 tweets every square kilometer. The density of the tweets at different land use categories revealed the usage frequency of twitter in these areas.

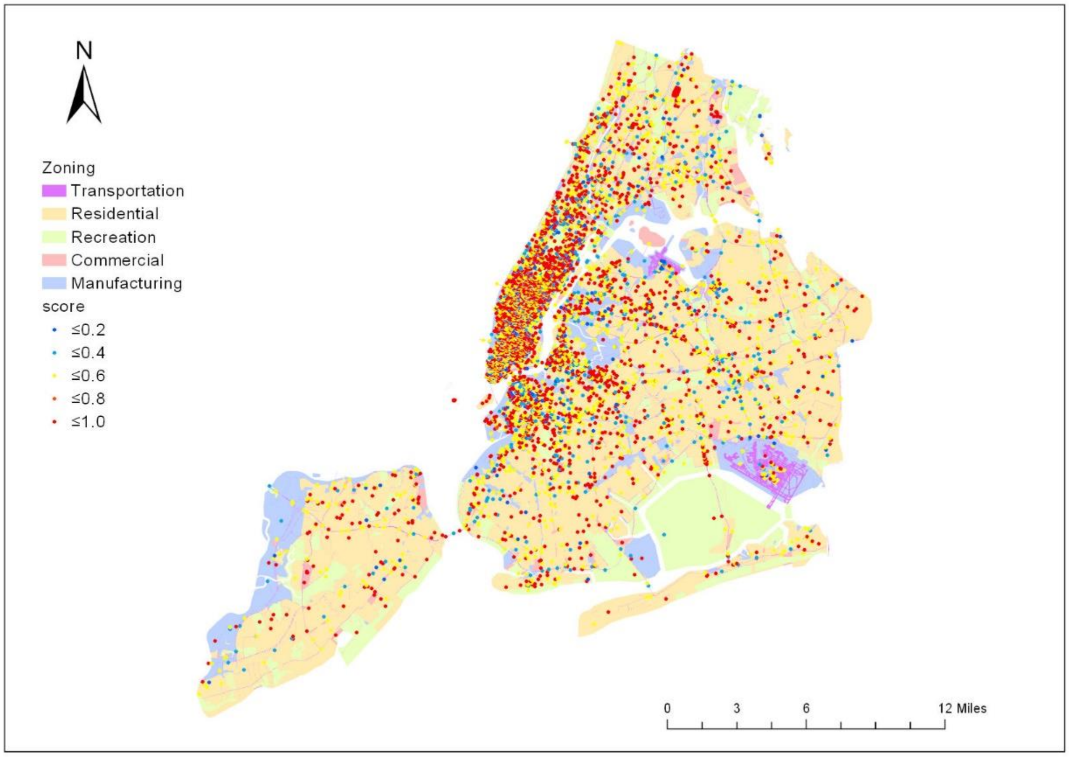


Figure 12 Distribution of the tweets and the land use categories

Table 3 Density of the tweets in different land use categories

ZONEDIST	Tweets	Area (km^2)	Density (tweets/ km^2)
Commercial	17508	41.1713	425.2477
Manufacturing	6652	114.2605	58.21784
Recreation	3309	141.8695	23.32425
Residential	12564	533.0148	23.57158
Transportation	3660	29.8058	122.7949

The annual INRIX Global Traffic Scorecard in 2018 indicates that New York City ranked fourth of the most congested city in the United States. Traffic problems can lead to aggressive driving even road rage.

Figure 10 shows the average sentiment score in the five land use categories. Generally, the sentiment appeared to be positive. The lowest average sentiment score was found in the transportation area with 0.1562 lower than the overall average of the sentiment score. The average sentiment score in recreation areas is the highest among the other land use categories with 0.0686 higher than the overall average.

The annual INRIX Global Traffic Scorecard in 2018 indicates that New York City ranked fourth of the most congested city in the United States. Traffic problems can lead to aggressive driving even road rage.

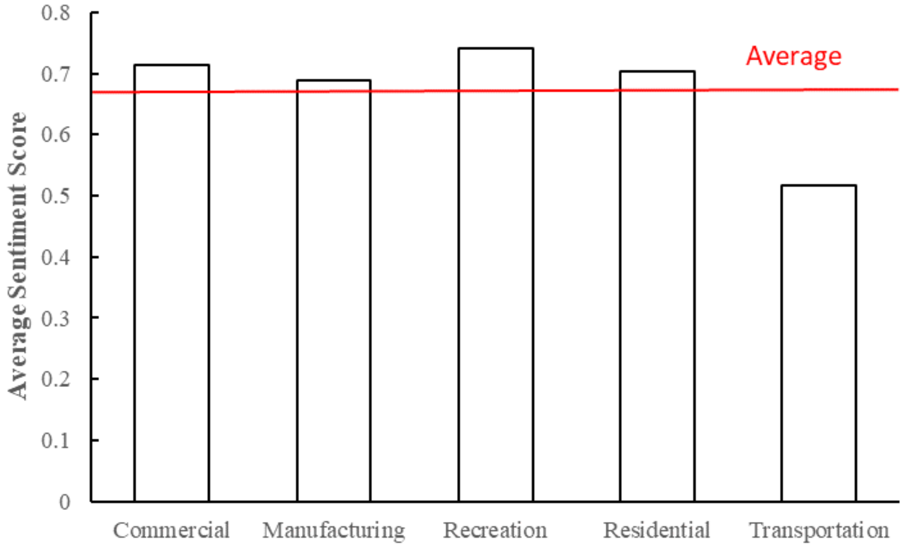


Figure 13 Average sentiment score by land use categories

Figure 11 of the average sentiment score within a week indicats that the public sentiment scores on weekdays are lower than the overall average while the sentiment scores at weekends are above the average. In a week, people’s good mood arrived at the peak on Saturday with score 0.0122 above the overall average rising from the trough on Tuesday 0.0193 below the average. On Monday, the sentiment fell below the average since people return to work and routine after break at weekends, which is known as Monday Blues. The emotion continued to fall till Tuesday and went up to the peak on Saturday.

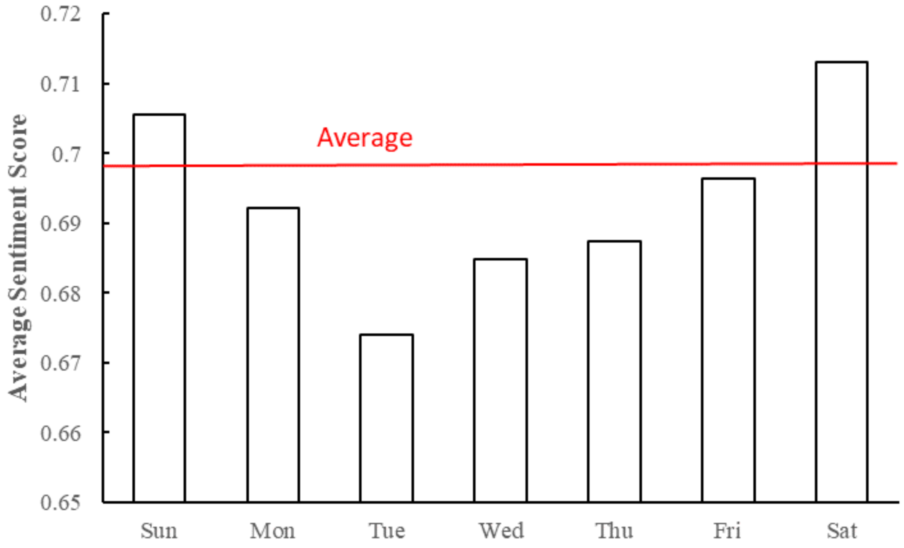


Figure 14 Average sentiment score by days of a week

The average sentiment score by time period of a day revealed that the sentiment score within a day remained high in the day and decreased at night with the lowest score before dawn, 0.0479 below the overall average. The public sentiment rose with a high speed above the average to the morning. Then the top at noon with 0.0222 higher than the overall average sentiment score.

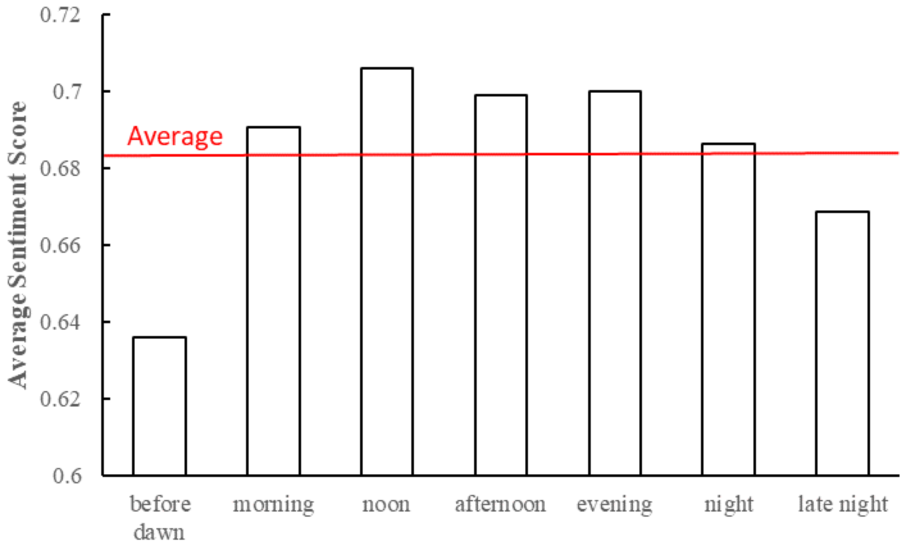


Figure 15 Average sentiment score by time period of a day

The spatiotemporal variances of the sentiment score are shown in Table 4. The matrix of days of a week and time period indicated that the overall sentiment was positive. Mood change occurred before dawn on Monday and Saturday. The matrix of land use categories and time periods revealed that the sentiment at the transportation area was lower than the other land use type all the time. Within a day, the sentiment at night time was lower than the day time.

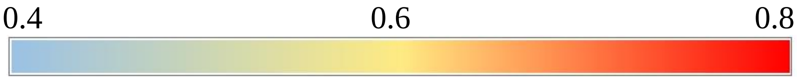


Table 4 Spatiotemporal variances of the sentiment score

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
before dawn	0.6798	0.6163	0.6326	0.6360	0.6310	0.6652	0.6065
morning	0.6924	0.7137	0.6883	0.6887	0.6735	0.6990	0.6853
noon	0.7340	0.6999	0.7093	0.6905	0.6924	0.6895	0.7366
afternoon	0.6870	0.6944	0.6816	0.7082	0.7015	0.7066	0.7258
evening	0.6923	0.6854	0.6778	0.6974	0.7075	0.7128	0.7378
night	0.7463	0.6951	0.6394	0.6728	0.6650	0.6909	0.7224
late night	0.7320	0.6675	0.6505	0.6323	0.6875	0.6358	0.6819

(a) days of a week and time periods

	Commercial	Manufacturing	Recreation	Residential	Transportation
before dawn	0.6537	0.7028	0.7468	0.6634	0.5261
morning	0.7127	0.6991	0.7637	0.6852	0.5319
noon	0.7198	0.7114	0.7312	0.7198	0.5120
afternoon	0.7124	0.6793	0.7391	0.7114	0.5535
evening	0.7299	0.6679	0.7435	0.7119	0.5011
night	0.7094	0.6735	0.7482	0.7053	0.4720
late night	0.6800	0.7303	0.6935	0.6792	0.4733

(b) land use categories and time periods

	Commercial	Manufacturing	Recreation	Residential	Transportation
Monday	0.7073	0.6622	0.7383	0.7209	0.4966
Tuesday	0.7054	0.6528	0.7522	0.6734	0.5286
Wednesday	0.7172	0.6911	0.7057	0.6831	0.4970
Thursday	0.7143	0.6719	0.7306	0.7063	0.5353
Friday	0.7231	0.7019	0.7293	0.6979	0.5118

Saturday	0.7205	0.7555	0.7722	0.7133	0.5165
Sunday	0.7140	0.7185	0.7408	0.7243	0.5169

(c) land use categories and days of a week

Effect of Land Use and Time

The Likelihood Ratio Test (Table 4) shown that the p value of the fixed effect model compared with the model without certain fixed effect revealed more than 99% likelihood that the effect of time period, days of the week and land use type are significant on the public sentiment. Among these fixed effects, the effect of land use categories is the largest. However, since the marginal R-squared of the model is 0.4905% and the conditional R-squared is 24.1212% the model cannot be used for prediction.

Table 5 Likelihood Ratio Test Results of the model

Fixed Effect	F value	P value
Weekday	3.1141	0.0047368**
Time Period	4.1745	0.0003356***
Zone	39.0789	<0.0001***

The fixed effect results from the multivariate linear mixed-effect regression model was shown in Table 5. The highest average sentiment score of each category was selected as the reference level of each category. The p values shown that the effect of land use categories was significant on the sentiment but not all of the fixed effect of the time period was significant. As was shown in the coefficients of different land use areas, the recreation earned most positive sentiment and the transportation areas had lowest sentiment compared with the others. The coefficients of the fixed effect reflected how much the sentiment decrease from the highest average sentiment score to the land use type or time period. For example, the sentiment decreased by

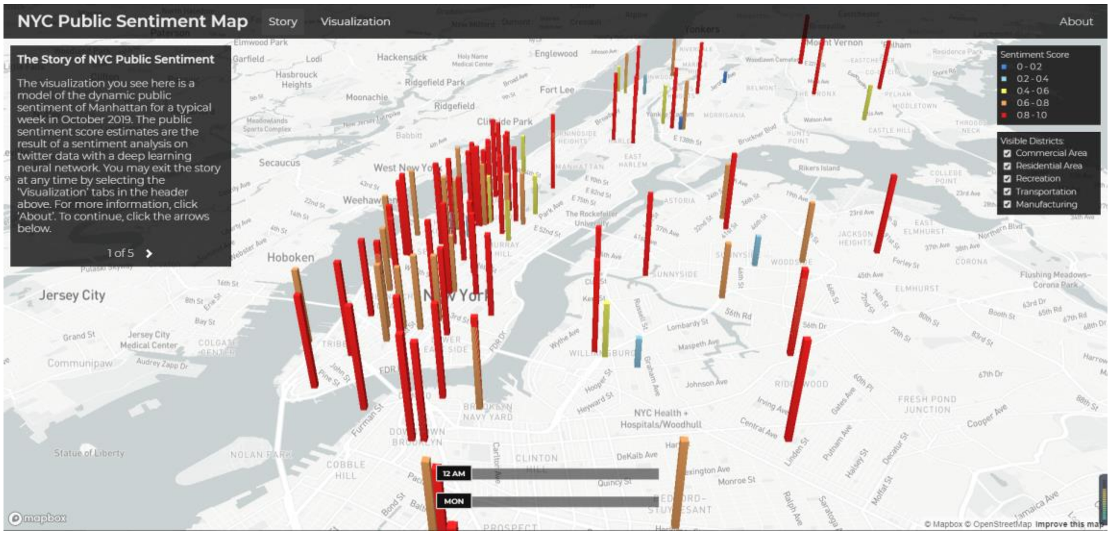
0.084362 in the transportation area (-0.040836) during late night (-0.028187) on Friday (-0.008280) than the sentiment in the recreation area before dawn on Saturday.

Table 6 Results of the coefficient of the fixed effect of the model

	Coefficients	Standard Error	P value
(Intercept)	0.763038	0.003848	<0.0001 ***
Monday	-0.005654	0.003636	0.02927 *
Tuesday	-0.006000	0.003689	0.03454 *
Wednesday	-0.009768	0.004072	0.09436
Thursday	-0.015339	0.003901	0.13138
Friday	-0.008280	0.003879	0.01865 *
Saturday	0.000000 (Reference)		
Sunday	-0.009480	0.003884	0.10035
Before Dawn	0.000000 (Reference)		
Morning	-0.011752	0.003288	0.09189 *
Noon	-0.010152	0.003148	0.05388
Afternoon	-0.012086	0.005130	0.00783 **
Evening	-0.009862	0.002998	0.04581 *
Night	-0.015913	0.003606	0.2886
Late Night	-0.028187	0.004545	0.00170 **
Manufacturing	-0.032706	0.003519	<0.0001 ***
Recreation	0.000000 (Reference)		
Residential	-0.009200	0.002643	0.00273 **
Transportation	-0.040836	0.004403	<0.0001 ***
Commercial	-0.001279	0.004031	0.007714 **

Visualization

The visualization is a model of the dynamic public sentiment of New York City for a typical week in October 2019 (Figure 13). The public sentiment score estimates are the result of a sentiment analysis on twitter data with a deep learning neural network. Users may exit the story at any time by selecting the ‘Visualization’ tabs in the header above. For more information, click ‘About’. To continue reading the story of different scenarios, click the arrows.



The Story of NYC Public Sentiment

The visualization you see here is a model of the dynamic public sentiment of Manhattan for a typical week in October 2019. The public sentiment score estimates are the result of a sentiment analysis on twitter data with a deep learning neural network. You may exit the story at any time by selecting the 'Visualization' tabs in the header above. For more information, click 'About'. To continue, click the arrows below.

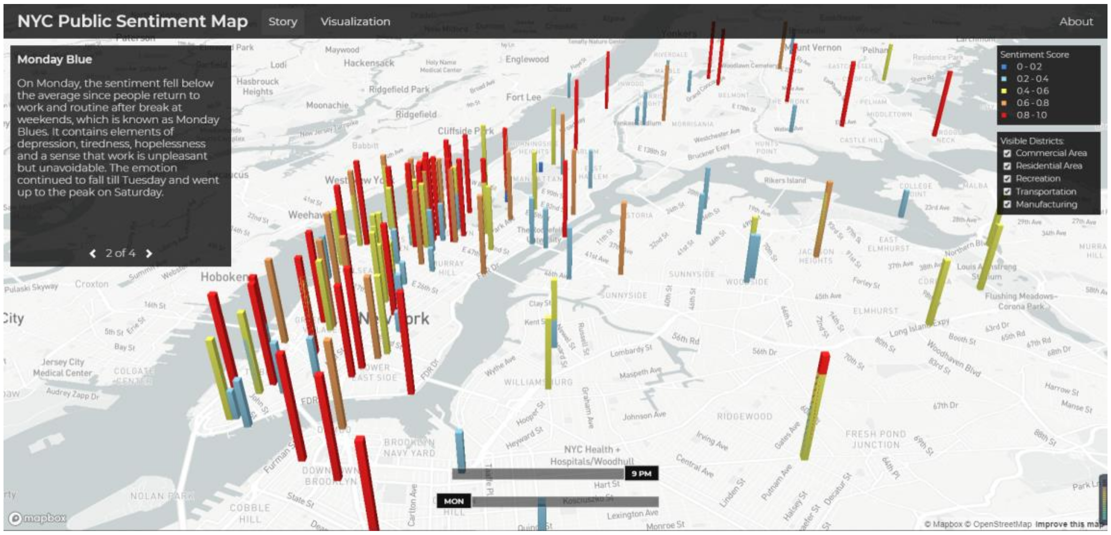
1 of 5 ›

Sentiment Score
- 🟦 0 - 0.2
- 🟦 0.2 - 0.4
- 🟨 0.4 - 0.6
- 🟧 0.6 - 0.8
- 🟥 0.8 - 1.0

Visible Districts:
- ☑ Commercial Area
- ☑ Residential Area
- ☑ Recreation
- ☑ Transportation
- ☑ Manufacturing

Figure 16 Introduction of the NYC Public Sentiment Map

The first scenario was about Monday Blue at 9 p.m. Monday (Figure 14). On Monday, the sentiment fell below the average since people return to work and routine after break at weekends, which is known as Monday Blues. It contains elements of depression, tiredness, hopelessness and a sense that work is unpleasant but unavoidable. The emotion continued to fall till Tuesday and went up to the peak on Saturday.

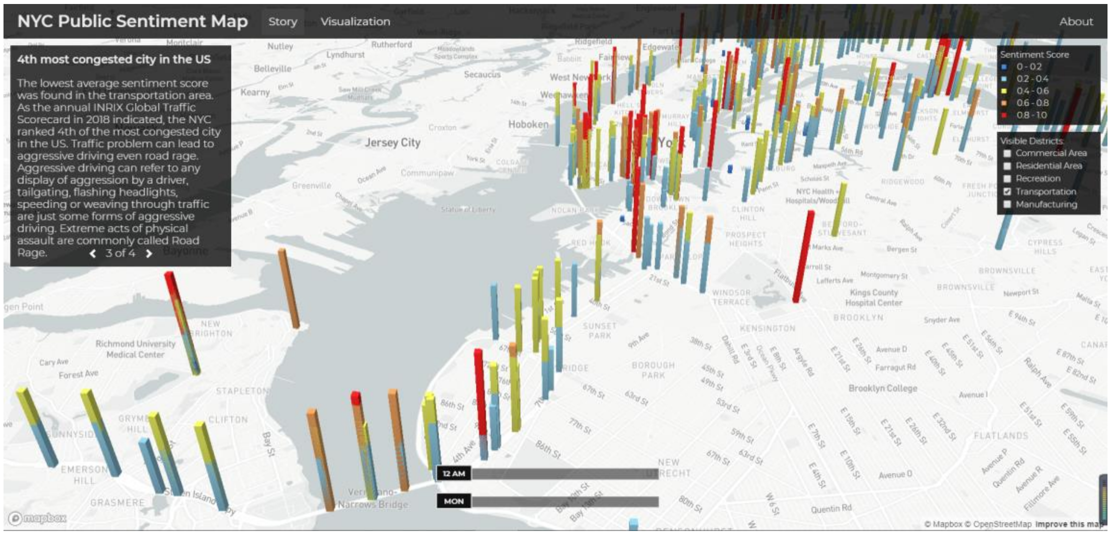


Monday Blue

On Monday, the sentiment fell below the average since people return to work and routine after break at weekends, which is known as Monday Blues. It contains elements of depression, tiredness, hopelessness and a sense that work is unpleasant but unavoidable. The emotion continued to fall till Tuesday and went up to the peak on Saturday.

Figure 17 Monday Blue scenario of the map

The second scenario discussed the low sentiment in transportation areas (Figure 15). The lowest average sentiment score was found in the transportation area. As the annual INRIX Global Traffic Scorecard in 2018 indicated, the NYC ranked 4th of the most congested city in the US. Traffic problems can lead to aggressive driving even road rage. Aggressive driving can refer to any display of aggression by a driver, tailgating, flashing headlights, speeding or weaving through traffic are just some forms of aggressive driving. Extreme acts of physical assault are commonly called Road Rage.

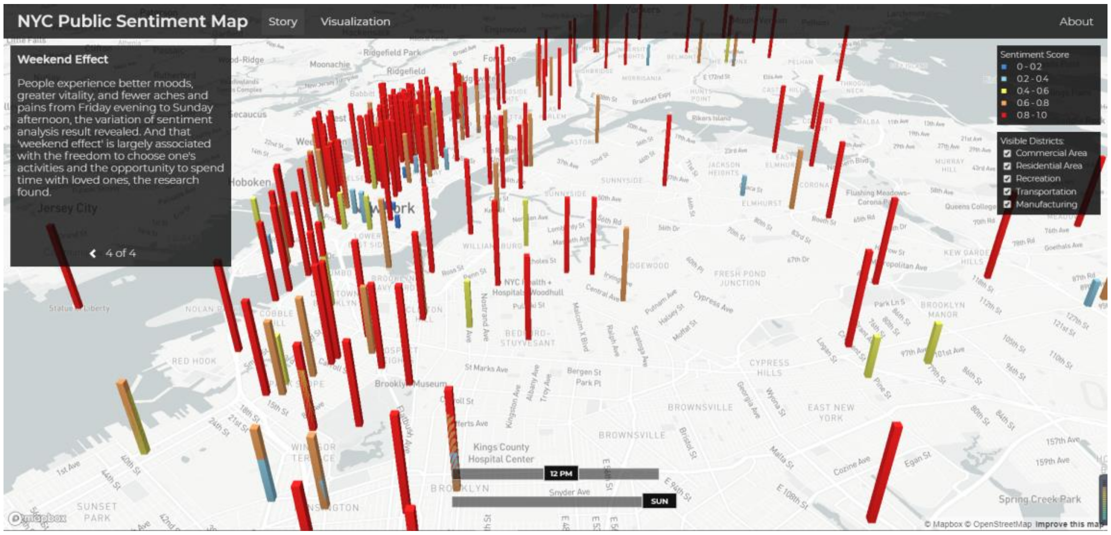


4th most congested city in the US

The lowest average sentiment score was found in the transportation area. As the annual INRIX Global Traffic Scorecard in 2018 indicated, the NYC ranked 4th of the most congested city in the US. Traffic problem can lead to aggressive driving even road rage. Aggressive driving can refer to any display of aggression by a driver, tailgating, flashing headlights, speeding or weaving through traffic are just some forms of aggressive driving. Extreme acts of physical assault are commonly called Road Rage.

Figure 18 Transportation scenario of the map

The last scenario indicated the weekend effect at 12 p.m. Sunday (Figure 16). People experience better moods, greater vitality, and fewer aches and pains from Friday evening to Sunday afternoon, the variation of sentiment analysis result revealed. And that 'weekend effect' is largely associated with the freedom to choose one's activities and the opportunity to spend time with loved ones, the research found.



Weekend Effect

People experience better moods, greater vitality, and fewer aches and pains from Friday evening to Sunday afternoon, the variation of sentiment analysis result revealed. And that 'weekend effect' is largely associated with the freedom to choose one's activities and the opportunity to spend time with loved ones, the research found.

Figure 19 Weekend Effect scenario of the map

At the visualization mode (Figure 17), the story content was hidden with more view on the data. Users can change the slider to see sentiment change at different time periods and see sentiment of different land use areas by checking the box of different land use categories.

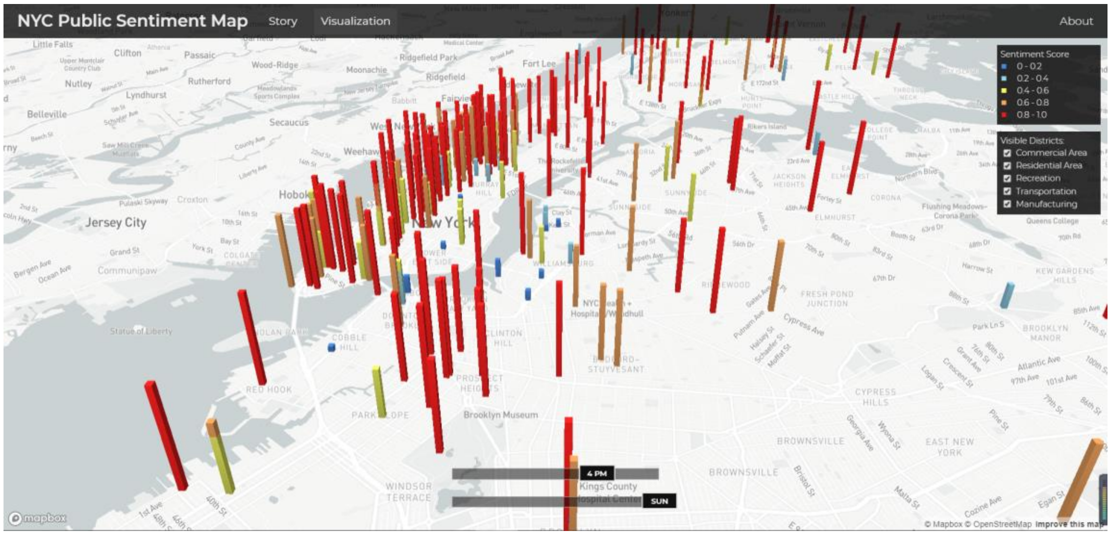
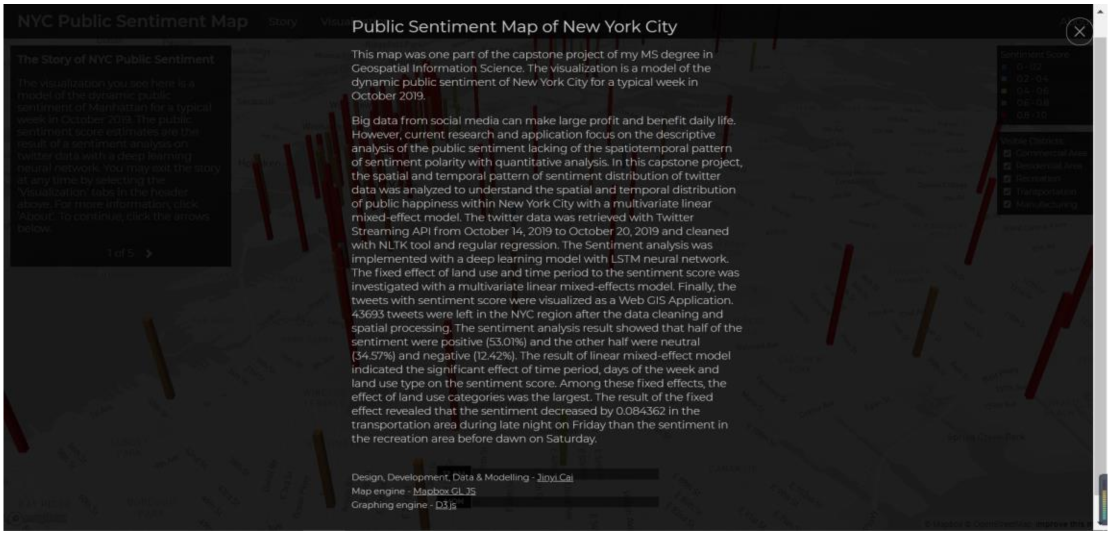


Figure 20 Visualization mode of the map

At the about page, the user can know about the background, data collection, data processing methods and analysis results of the map application. Besides, the information of the developer, map engine and graphing engine were shown at the bottom.



The Story of NYC Public Sentiment

The visualization you see here is a model of the dynamic public sentiment of Manhattan for a typical week in October 2019. The public sentiment score estimates are the result of a sentiment analysis on twitter data with a deep learning neural network. You may exit the story at any time by selecting the 'Visualization' tabs in the header above. For more information, click 'About'. To continue, click the arrows below.

1 of 5 ›

Public Sentiment Map of New York City

This map was one part of the capstone project of my MS degree in Geospatial Information Science. The visualization is a model of the dynamic public sentiment of New York City for a typical week in October 2019.

Big data from social media can make large profit and benefit daily life. However, current research and application focus on the descriptive analysis of the public sentiment lacking of the spatiotemporal pattern of sentiment polarity with quantitative analysis. In this capstone project, the spatial and temporal pattern of sentiment distribution of twitter data was analyzed to understand the spatial and temporal distribution of public happiness within New York City with a multivariate linear mixed-effect model. The twitter data was retrieved with Twitter Streaming API from October 14, 2019 to October 20, 2019 and cleaned with NLTK tool and regular regression. The Sentiment analysis was implemented with a deep learning model with LSTM neural network. The fixed effect of land use and time period to the sentiment score was investigated with a multivariate linear mixed-effects model. Finally, the tweets with sentiment score were visualized as a Web GIS Application. 43693 tweets were left in the NYC region after the data cleaning and spatial processing. The sentiment analysis result showed that half of the sentiment were positive (53.01%) and the other half were neutral (34.57%) and negative (12.42%). The result of linear mixed-effect model indicated the significant effect of time period, days of the week and land use type on the sentiment score. Among these fixed effects, the effect of land use categories was the largest. The result of the fixed effect revealed that the sentiment decreased by 0.084362 in the transportation area during late night on Friday than the sentiment in the recreation area before dawn on Saturday.

Design, Development, Data & Modelling - Jinyi Cai
Map engine - Mapbox GL JS
Graphing engine - D3.js

Figure 21 About page of the map

Chapter 5: Future Work and Conclusion

Conclusions

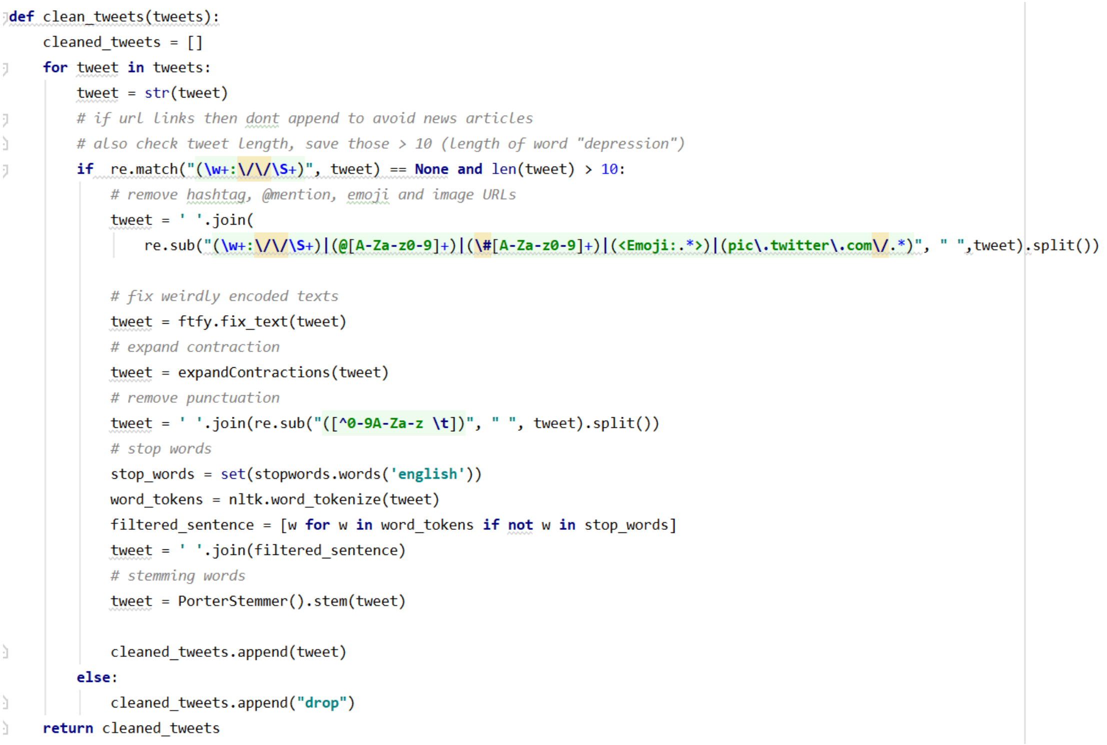
The study analyzed and visualized the spatial and temporal pattern of sentiment distribution of twitter data within a week to understand the spatial and temporal distribution of public happiness within New York City with a multivariate linear mixed-effect model. The sentiment analysis result showed that half of the sentiment was positive (53.01%) and the other half were neutral (34.57%) and negative (12.42%). The result of linear mixed-effect model indicated the significant effect of time period, days of the week and land use type on the sentiment score. Among these fixed effects, the effect of land use categories is the largest. The result of the fixed effect revealed that the sentiment decreased by 0.084362 in the transportation area (-0.040836) during late night (-0.028187) on Friday (-0.008280) than the sentiment in the recreation area before dawn on Saturday.

Future Work

There are several limitations to this study. In the first place, the bias existed on the representative of the population since only 20.1% of the Internet users used twitter platform according to the report of eMarketer (eMarketer, 2018). Besides, the georeferenced tweets just accounted for 1% of the whole tweets (Morstatter, Pfeffer, Liu, & Carley, 2013). Therefore, the population of twitter users could be different from the actual population, which leads to limitation to the sentiment analysis. For further study, one of the solutions for this bias is to integrate different source data. By combining data with different biases, the analysis will be more comprehensive. To get around the demographic limitation of the tweets, the probable gender and age can be inferred with the combination of different survey source (Lansley & Longley, 2016). For the sentiment analysis, the accuracy of the model was not much ideal. The model can be improved by tuning the hyperparameters and adding more hidden layers. In addition, the complication of sentiment does not restrict to three categories. Human emotions can be further classified into eight types: anger, fear, joy, sadness, disgust, surprise, trust, and anticipation (Plutchik,1980). The sentiment analysis model can be further to recognize more sentiment categories. For the regression analysis, the result showed that not all of the fixed effect was significant which could be influenced by the data bias, the accuracy of the sentiment analysis and the selection of fixed effect. For this reason, in addition to the improvement of the preview discussion, other fixed effects such as distance to park, household income and employment rate can be tested to be the variables of the regression model.

Appendix A

Data cleaning

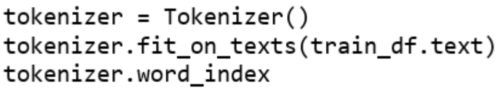


```
def clean_tweets(tweets):
    cleaned_tweets = []
    for tweet in tweets:
        tweet = str(tweet)
        # if url links then dont append to avoid news articles
        # also check tweet length, save those > 10 (length of word "depression")
        if  re.match("(\w+:\/\/\S+)", tweet) == None and len(tweet) > 10:
            # remove hashtag, @mention, emoji and image URLs
            tweet = ' '.join(
                re.sub("(\w+:\/\/\S+)|(@[A-Za-z0-9]+)|(\#[A-Za-z0-9]+)|(<Emoji:.*>)|(pic\.twitter\.com\/.*)", " ",tweet).split())

            # fix weirdly encoded texts
            tweet = ftfy.fix_text(tweet)
            # expand contraction
            tweet = expandContractions(tweet)
            # remove punctuation
            tweet = ' '.join(re.sub("([^0-9A-Za-z \t])", " ", tweet).split())
            # stop words
            stop_words = set(stopwords.words('english'))
            word_tokens = nltk.word_tokenize(tweet)
            filtered_sentence = [w for w in word_tokens if not w in stop_words]
            tweet = ' '.join(filtered_sentence)
            # stemming words
            tweet = PorterStemmer().stem(tweet)

            cleaned_tweets.append(tweet)
        else:
            cleaned_tweets.append("drop")
    return cleaned_tweets
```

Figure 22 Data cleaning

Sentiment analysis

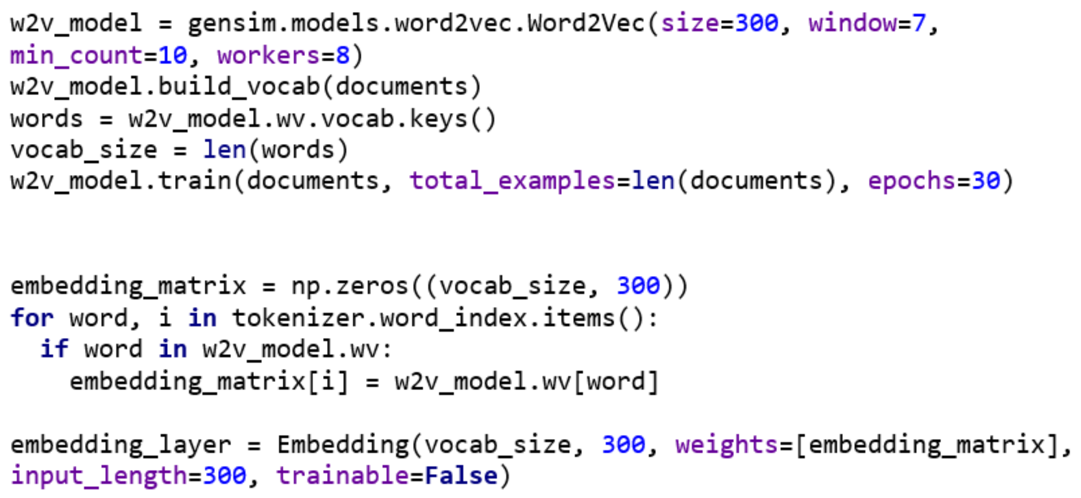
Tokenizer



```
tokenizer = Tokenizer()
tokenizer.fit_on_texts(train_df.text)
tokenizer.word_index
```

Figure 23 Tokenizer

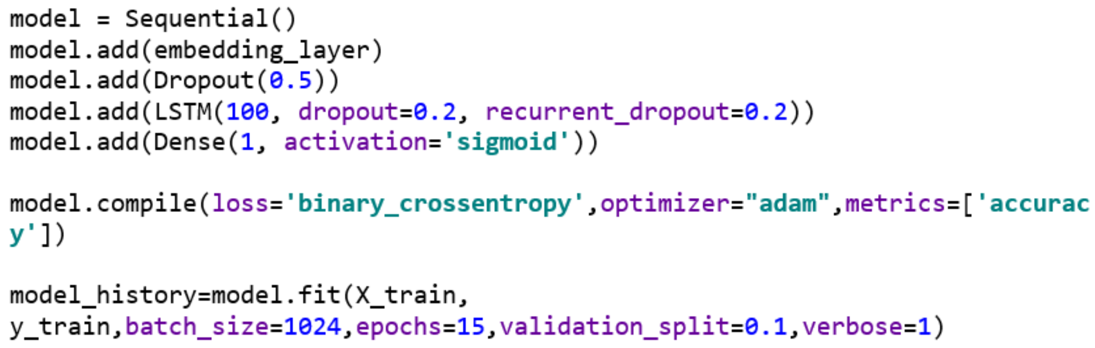
Set up the embedding layer



```
w2v_model = gensim.models.word2vec.Word2Vec(size=300, window=7,
min_count=10, workers=8)
w2v_model.build_vocab(documents)
words = w2v_model.wv.vocab.keys()
vocab_size = len(words)
w2v_model.train(documents, total_examples=len(documents), epochs=30)


embedding_matrix = np.zeros((vocab_size, 300))
for word, i in tokenizer.word_index.items():
  if word in w2v_model.wv:
    embedding_matrix[i] = w2v_model.wv[word]

embedding_layer = Embedding(vocab_size, 300, weights=[embedding_matrix],
input_length=300, trainable=False)
```

Figure 24 Embedding layer

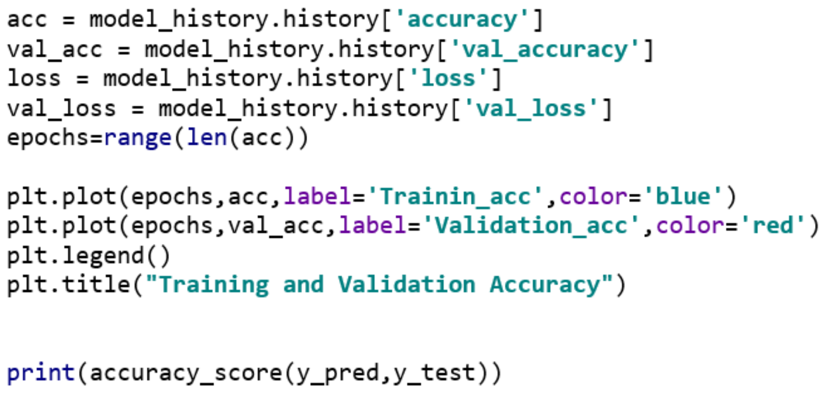
Building the classification model



```
model = Sequential()
model.add(embedding_layer)
model.add(Dropout(0.5))
model.add(LSTM(100, dropout=0.2, recurrent_dropout=0.2))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary_crossentropy',optimizer="adam",metrics=['accurac
y'])

model_history=model.fit(X_train,
y_train,batch_size=1024,epochs=15,validation_split=0.1,verbose=1)
```

Figure 25 classification model

Accuracy assessment

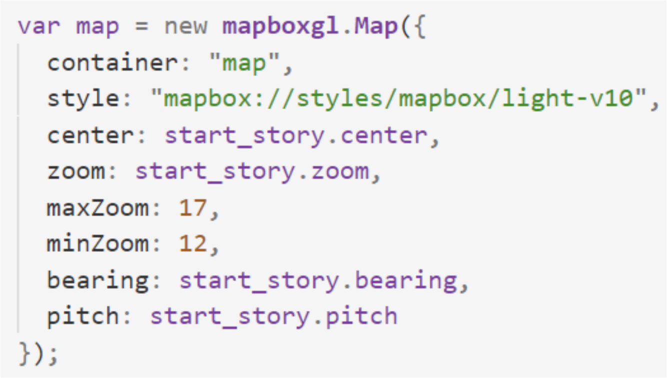


```
acc = model_history.history['accuracy']
val_acc = model_history.history['val_accuracy']
loss = model_history.history['loss']
val_loss = model_history.history['val_loss']
epochs=range(len(acc))

plt.plot(epochs,acc,label='Trainin_acc',color='blue')
plt.plot(epochs,val_acc,label='Validation_acc',color='red')
plt.legend()
plt.title("Training and Validation Accuracy")


print(accuracy_score(y_pred,y_test))
```

Figure 26 Accuracy assessment

<u>Visualization:</u>

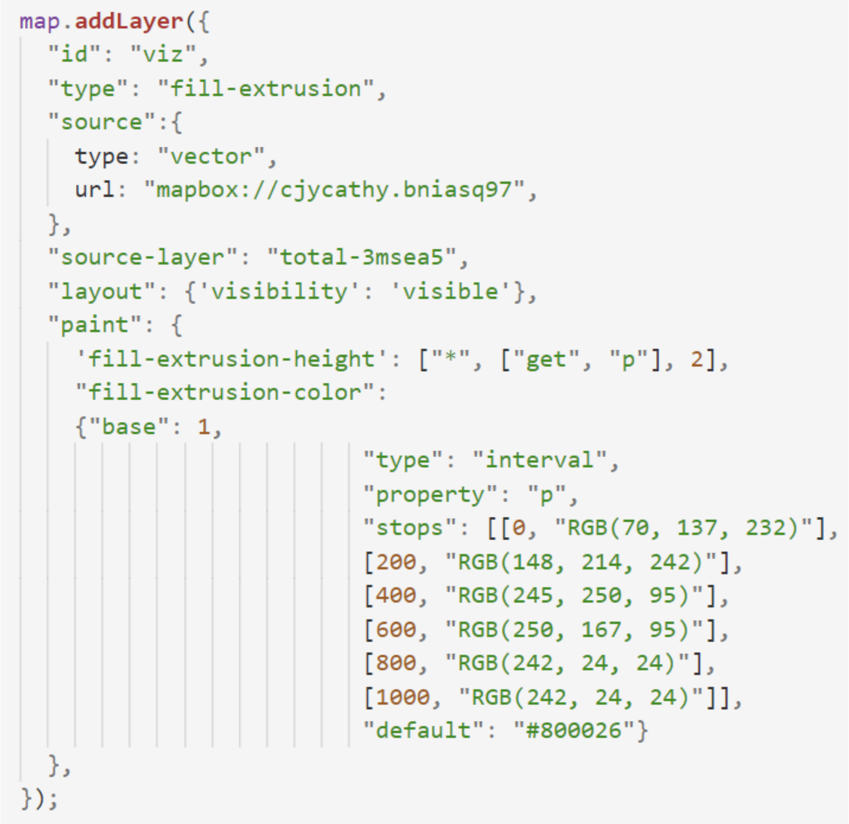
Use the new mapboxgl.Map to create a new map object.



```
var map = new mapboxgl.Map({
  container: "map",
  style: "mapbox://styles/mapbox/light-v10",
  center: start_story.center,
  zoom: start_story.zoom,
  maxZoom: 17,
  minZoom: 12,
  bearing: start_story.bearing,
  pitch: start_story.pitch
});
```

Figure 27 mapboxgl.Map Object

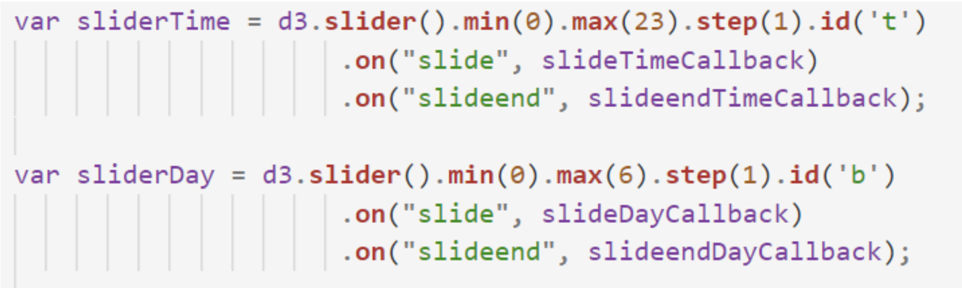
Use the addlayer() method to add a Mapbox style layer to the map's style with the tileset source uploaded to the Mapbox studio.



```
map.addLayer({
  "id": "viz",
  "type": "fill-extrusion",
  "source":{
    type: "vector",
    url: "mapbox://cjycathy.bniasq97",
  },
  "source-layer": "total-3msea5",
  "layout": {'visibility': 'visible'},
  "paint": {
    'fill-extrusion-height': ["*", ["get", "p"], 2],
    "fill-extrusion-color":
    {"base": 1,
                      "type": "interval",
                      "property": "p",
                      "stops": [[0, "RGB(70, 137, 232)"],
                      [200, "RGB(148, 214, 242)"],
                      [400, "RGB(245, 250, 95)"],
                      [600, "RGB(250, 167, 95)"],
                      [800, "RGB(242, 24, 24)"],
                      [1000, "RGB(242, 24, 24)"]],
                      "default": "#800026"}
  },
});
```

Figure 28 addlayer() method

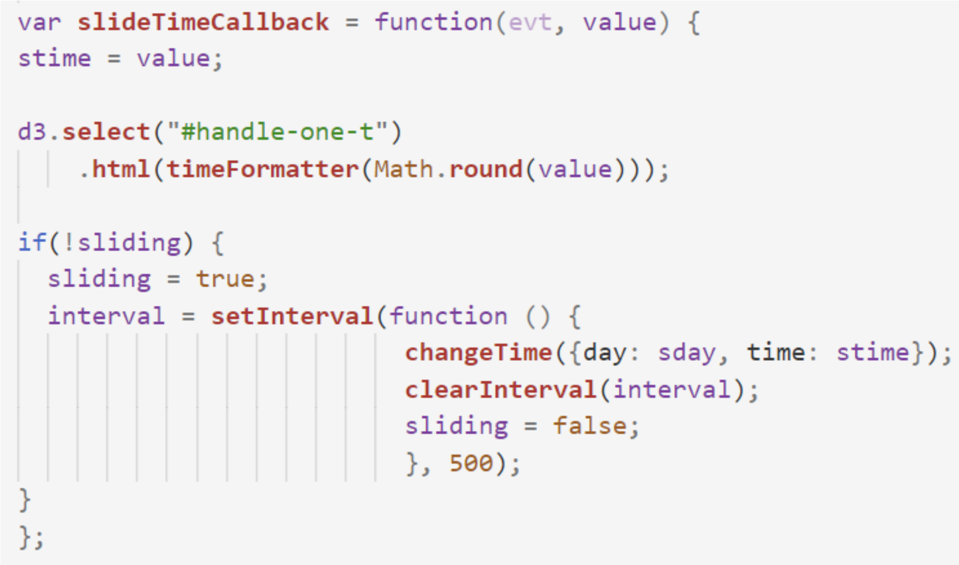
The time range within the days of the week and hours of the day was implemented by drawing a slider with d3.slider() .



```
var sliderTime = d3.slider().min(0).max(23).step(1).id('t')
                     .on("slide", slideTimeCallback)
                     .on("slideend", slideendTimeCallback);

var sliderDay = d3.slider().min(0).max(6).step(1).id('b')
                     .on("slide", slideDayCallback)
                     .on("slideend", slideendDayCallback);
```

Figure 29 d3.slider() method

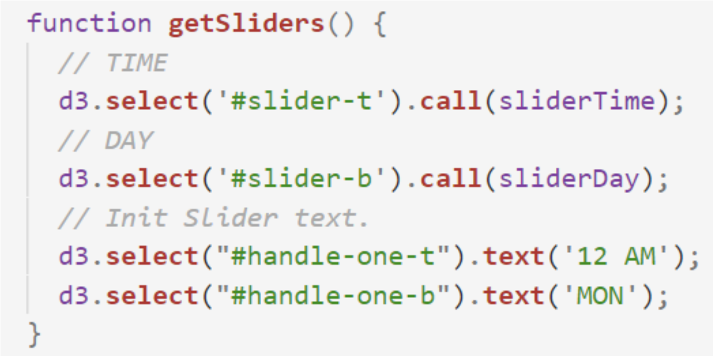
Use the slideTimeCallback() and slideendTimeCallback() function to return the time range to show the time change on the page and get the sentiment score..



```
var slideTimeCallback = function(evt, value) {
stime = value;

d3.select("#handle-one-t")
    .html(timeFormatter(Math.round(value)));

if(!sliding) {
  sliding = true;
  interval = setInterval(function () {
                        changeTime({day: sday, time: stime});
                        clearInterval(interval);
                        sliding = false;
                        }, 500);
}
};
```

Figure 30 slideTimeCallback() and slideendTimeCallback() function

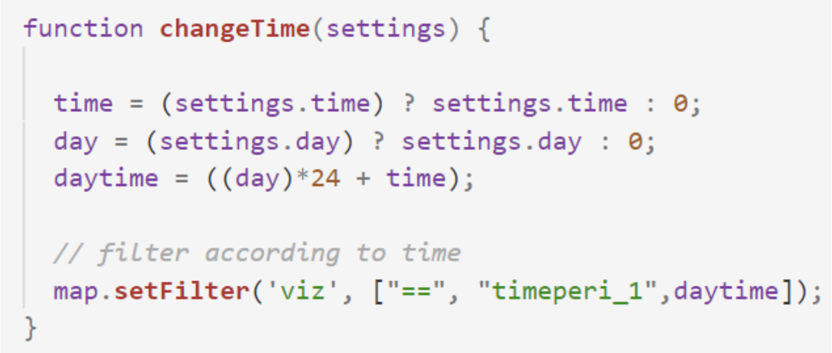
Use the d3.select() method to call the d3.slider() method



```
function getSliders() {
  // TIME
  d3.select('#slider-t').call(sliderTime);
  // DAY
  d3.select('#slider-b').call(sliderDay);
  // Init Slider text.
  d3.select("#handle-one-t").text('12 AM');
  d3.select("#handle-one-b").text('MON');
}
```

Figure 31 d3.select() method

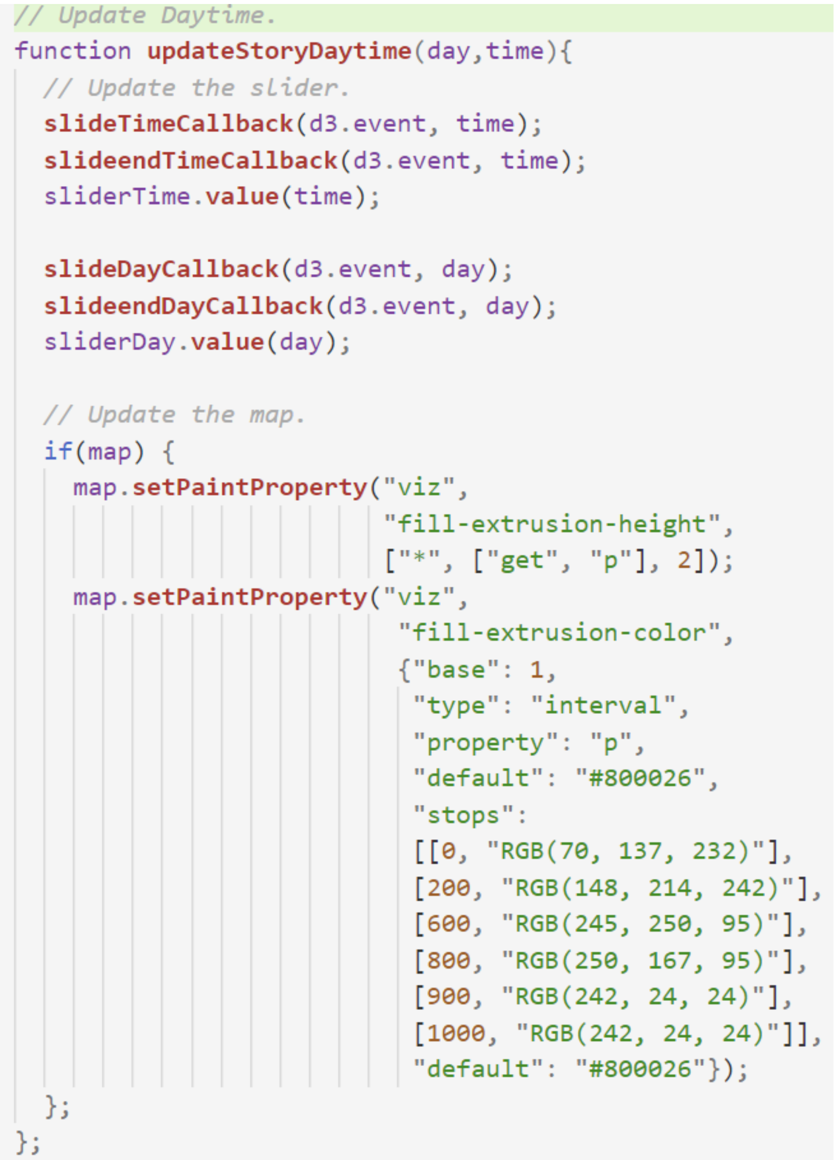
Use the changeTime() function to update the data shown on the map with the setFilter() method of the map object.



```
function changeTime(settings) {

  time = (settings.time) ? settings.time : 0;
  day = (settings.day) ? settings.day : 0;
  daytime = ((day)*24 + time);

  // filter according to time
  map.setFilter('viz', ["==", "timeperi_1",daytime]);
}
```

Figure 32 changeTime() function

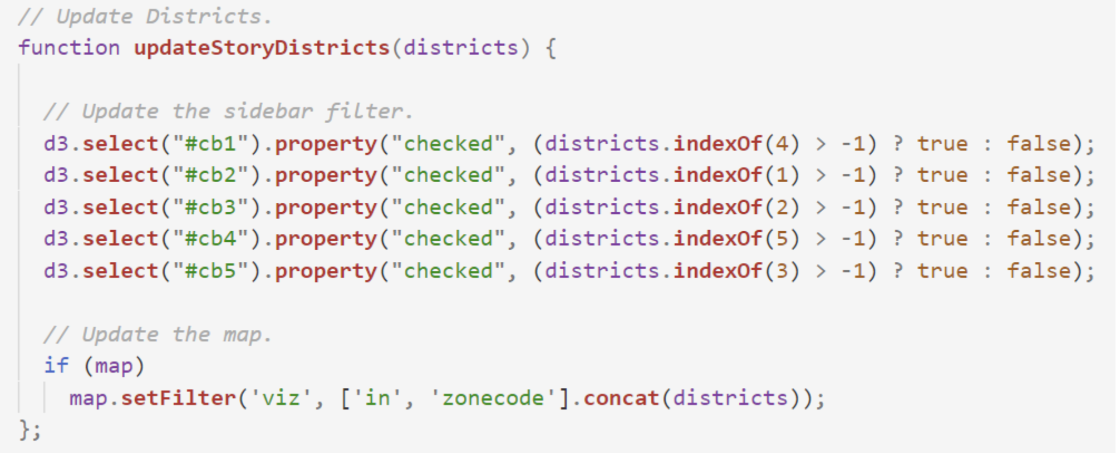
In the story mode, the slider changed with different story by a updateStoryDaytime() function, which called the slideTimeCallback() function to update the data shown on the map with the setFilter() method at nyc.js. The sliders on the map were moved by the value() method of the sliderTime object.



```
// Update Daytime.
function updateStoryDaytime(day,time){
  // Update the slider.
  slideTimeCallback(d3.event, time);
  slideendTimeCallback(d3.event, time);
  sliderTime.value(time);

  slideDayCallback(d3.event, day);
  slideendDayCallback(d3.event, day);
  sliderDay.value(day);

  // Update the map.
  if(map) {
    map.setPaintProperty("viz",
                          "fill-extrusion-height",
                          ["*", ["get", "p"], 2]);
    map.setPaintProperty("viz",
                           "fill-extrusion-color",
                           {"base": 1,
                            "type": "interval",
                            "property": "p",
                            "default": "#800026",
                            "stops":
                            [[0, "RGB(70, 137, 232)"],
                            [200, "RGB(148, 214, 242)"],
                            [600, "RGB(245, 250, 95)"],
                            [800, "RGB(250, 167, 95)"],
                            [900, "RGB(242, 24, 24)"],
                            [1000, "RGB(242, 24, 24)"]],
                            "default": "#800026"});
  };
};
```

Figure 33 updateStoryDaytime() function

Use the updateStoryDistricts() function with d3.select().property() method to check the district box as the story page change. The data on the map was updated with the setFilter() method .



```
// Update Districts.
function updateStoryDistricts(districts) {

  // Update the sidebar filter.
  d3.select("#cb1").property("checked", (districts.indexOf(4) > -1) ? true : false);
  d3.select("#cb2").property("checked", (districts.indexOf(1) > -1) ? true : false);
  d3.select("#cb3").property("checked", (districts.indexOf(2) > -1) ? true : false);
  d3.select("#cb4").property("checked", (districts.indexOf(5) > -1) ? true : false);
  d3.select("#cb5").property("checked", (districts.indexOf(3) > -1) ? true : false);

  // Update the map.
  if (map)
    map.setFilter('viz', ['in', 'zonecode'].concat(districts));
};
```

Figure 34 updateStoryDistricts() function

Bibliography

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Cao, X., Macnaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. G. (2018). Using twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, *15*(2). https://doi.org/10.3390/ijerph15020250

Eldering, R. (2017). *Thesis Report GIRS-2017-13 Using sentiment analysis on Twitter data to identify place experience in New York City , United States of America Using sentiment analysis on Twitter data to identify place experience in New York City , United States of America*.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 400–408.

Plunz, R. A., Zhou, Y., Carrasco Vintimilla, M. I., Mckeown, K., Yu, T., Uguccioni, L., & Sutto, M. P. (2019). Twitter sentiment in New York City parks as a measure of well-being. *Landscape and Urban Planning*, *189*, 235–246.

https://doi.org/10.1016/J.LANDURBPLAN.2019.04.024

Serrano-guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-viedma, E. (2015). Sentiment analysis : A review and comparative analysis of web services. *INFORMATION SCIENCES*, *311*, 18–38. https://doi.org/10.1016/j.ins.2015.03.040

Yang, W., & Mu, L. (2015). GIS analysis of depression among Twitter users. *Applied Geography*, *60*, 217–223. https://doi.org/10.1016/j.apgeog.2014.10.016

Song, Z and Xia, J. 2016. Spatial and Temporal Sentiment Analysis of Twitter data. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, Pp. 205–221. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.p. License: CC-BY 4.0.

Stojanovski, D, Chorbev, I, Dimitrovski, I and Madjarov, G. 2016. Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, Pp. 223–235. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.q. License: CC-BY 4.0.

eMarketer. (2018, March 15). Latest eMarketer Social Forecast Shows Shifting Platform Use. Retrieved November 3, 2019, from https://www.emarketer.com/content/latest-emarketer-social-forecast-shows-shifting-platform-use.

Plutchik, R. (1980). *Emotion: a psychoevolutionary synthesis*. New York, NY: Harper & Row.