Sentiment Analysis on Covid-19 Twitter Data

Amrutha Ragothaman and Ching-Yu Huang

***Abstract*—According to the WHO, Covid-19 is an illness caused by a novel coronavirus scientifically known as SARS-CoV2. This virus was first discovered in late December 2019, when a cluster of pneumonia cases was reported from Wuhan, China. Since then, the virus has continued to spread and cases have grown rapidly, leading to 74.3 million cases worldwide and 1.65 million deaths in within the span of a year.**

According to Dr. Sanil Rege, a psychiatrist, the effects of quarantining and social distancing include multiple stressors affect a person during a time of isolation. For instance, a person could have unpleasant experiences due to loss of freedom, separation from significant people in their lives, fear, financial stability, and lack of supplies. These factors can strongly lead to the development of stress symptoms such as irritability, insomnia, temper issues, emotional burnout, and overall low mental health. As such, the Covid-19 pandemic has affected the normalcy of life throughout the world and many people have taken to social media platforms such as Twitter to express their thoughts and feelings.

In order to understand the type of discussions taking place regarding Covid-19 and to recognize major topics of concern, the relationship between tweet sentiment and Covid-19 casualties are analyzed, then, Tweet text is examined to identify frequently used words, hashtags, and mentioned users. Covid-19 keyword containing tweets are downloaded using Tweepy to a database and analyzed for sentiment by NLTK Vader. Results suggest a moderate positive correlation between negative sentiment and Covid-19 cases and deaths.

***Index Terms*—COVID-19, Twitter, sentiment analysis, social media.**

I. Introduction

The pandemic has exerted a prominent influence on people's lives over the past year. Social media platforms provide an excellent base for communicating opinions. Of all platforms available, Twitter is a strong resource to get opinions as Tweets are posted constantly by users and can be downloaded live, allowing for real time data collection.

To conduct this research, the tweets are collected using a Twitter Streamer program using the Tweepy API available for Python [1], sentiment is then analyzed using NLTK Vader [2] following Beri's work as an example [3]. After obtaining sentiment, the compound sentiment values will be tested for correlation to Covid-19 data.

Morrissey *et al*. discuss the process of cleaning tweets and extracting textual data [4]. It is important to analyze the text of a tweet as its contents reveal more about the frequencies of words used, hashtags used, and users mentioned thereby enabling the identification of key concerns within the population.

This research intends to perform exploratory analysis on Tweet sentiments from tweets downloaded over a period of time to the number of Covid-19 cases and deaths from four large cities in the United States to establish a relationship between tweet sentiment and Covid-19 casualties.

Manuscript received January 6, 2021; revised April 10, 2021.

Amrutha Ragothaman was with School of Computer Science and Technology, Kean University, Union, NJ, 07083 USA (e-mail: ragothaa@kean.edu).

Ching-yu Huang is with the School of Computer Science & Technology, Kean University, Union, NJ 07083, USA (e-mail: chuang@kean.edu).

II. Materials

This research datasets are based on two sources: Twitter and Covid-19. The details of the datasets are described below.

A. Twitter Data

Twitter data was collected from August 18, 2020 to October 3, 2020 using Tweepy [1]. The program streamed only English tweets, extracted fields with key information from the tweet, and stored them in a 'tweets' table in the vader.kean.edu database.

This streamer used the following keywords to filter tweets: ["covid", "covid19", "corona", "coronavirus", "corona virus", "covid-19", "covid_19", "covid 19", "quarantine"]. The twitter streamer and cronjob were run on yoda.kean.edu Linux server.

Twitter data was collected for 2 hours a day; 10:00 AM EST to 11:00 AM EST and 10:00 PM EST and 11:00 PM EST from August 18th to September 2nd. Starting September 3rd till October 3rd, however, data was collected 24 hours a day. All tweets were stored in both text files with the whole JSON dictionary for each tweet as well as in the database.

There is a total of 73,592,078 tweets in the database. Within the timeframe of August 18th to September 2nd, number of tweets per day ranged between 47,567 – 300,694 and the average number of tweets per day is 217,802.9. For the timeframe of September 3rd to October 3rd, number of tweets per day ranged from 1,281,903 – 4,200,525 and the average number of tweets per day is 2,261,523.5. File sizes for this timeframe were larger as tweets were downloaded for 2 hours at once, 0.106GB – 0.327GB, as opposed to one hour twice a day ranging from 92KB – 138k KB.

To get an estimate of how many tweets are inserted to the database, a file, 2020-09-10-12_07_01.txt.gz of size 0.308GB is examined. Firstly, the number of tweets within that file is obtained iteratively – 363,552, then that count is compared with a query of the database to get count within the timeframe the tweets in the file were streamed for, from 12:07:01 to 14:07:01 – 353,243. This will help identify amount of data loss that has occurred – 10,309 tweets, 2.84% loss.

Additionally, due to an error during tweet streaming, the database is missing tweet data between the hours of 8:00 PM to 12:00 AM on September 24th and 12:00 AM to 4:00 AM,

DOI: 10.7763/IJCTE.2021.V13.1297

11:00 PM to 12:00 AM on September 25th and tweets on September 21st, 22nd, 23rd, and 25th are missing data regarding location. Therefore, the analysis will not use data obtained on these 4 days.

B. Covid-19

Covid-19 data was obtained from the New York Times opensource repository on GitHub [5] and was accessed using the io and requests packages and stored for analysis in a data frame using the Pandas [6] package in Python. All analyses of the data were done in Python.

C. Data Processing and Selection

As the 'tweets' table was too large, the analysis was limited to 4 select cities filtered by the place column in the tweets table. The 4 cities are Los Angeles, New York, Houston, and Chicago.

Data for each city was collected by zip code for Houston [7], by principal cities in the metro area in Los Angeles [8], by the 5 boroughs for New York, and regions within the metro area inside Chicago city [9].

Covid-19 data for each city was obtained from New York Times county-wise Covid-19 dataset [5]. Within the dataset, New York data for all boroughs and the Richmond counties are classified as New York city, Los Angeles city uses Los Angeles county Covid-19 data, Houston data is obtained using the same counties used for filtering tweets [7], and counties within the metro area is used to obtain the Covid-19 data for Chicago [10].

A total of 30,508 tweets from Table I were used for this analysis (0.041% percent of the total dataset). Of all cities, New York had the greatest number of tweets – 12,790, followed by Los Angeles – 9738. Chicago and Houston have much fewer tweets than the other two cities at 3965 and 4015.

TABLE I: TWEET STRUCTURE AND EXAMPLE OF DATA IN COLUMNS USED

created_at	twt_text	place
2020-08-18 00:58:09	@RodentWild @Charlott...	Brooklyn, NY!US
2020-08-18 00:58:13	1. Movie 2. Karaoke 3. ...	Damansara, Selangor!MY
2020-08-18 00:58:21	40 percent of this count...	South Carolina, USA!US
2020-08-18 00:58:24	#TrumpCanceledAmeric...	West Virginia, USA!US
2020-08-18 00:58:25	@ApieD36 @NASHUASC...	New Hampshire, USA!US

III. METHODS

Firstly, the tweets are queried into Python using MySQL connector as a Pandas data frame. For analyzing tweet sentiment, only the created at and tweet text columns from the data are used. NLTK Vader [2] is used to obtain sentiment scores for each tweet in the query.

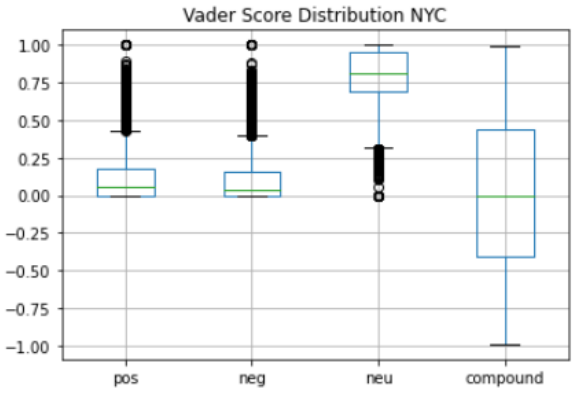
Vader returns a JSON dictionary classifying sentiment into 4 categories:

- Pos – the percent positivity of the text
- Neg – the percent negativity of the text
- Neu – percent neutrality of the text
- Compound – the overall aggregated score of the text

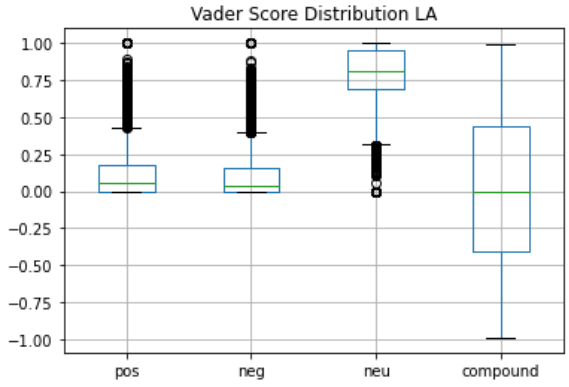
This study uses the compound score to conduct an analysis of Tweet sentiment as it numerically provides a categorization of the text sentiment. Text with a compound score ranging between -1 to +1 indicates the sentiment as negative or positive and tweets between the range of -0.05 to +0.05 are considered to be neutral [11].

A. Box Plot Analysis for Outliers

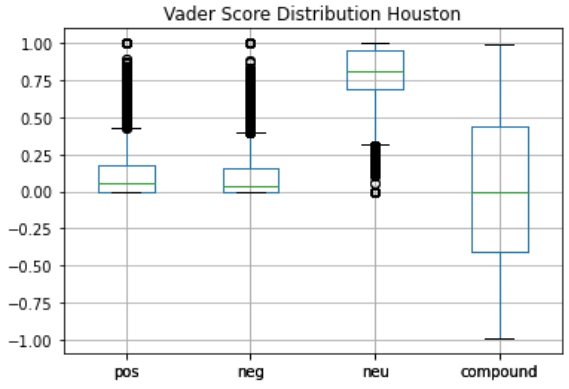
After obtaining the compound scores for all tweets, box plots are drawn to detect the presence of outliers for each city as shown in Fig. 1. The compound scores for all 4 cities suggest the distribution of compound scores is even.



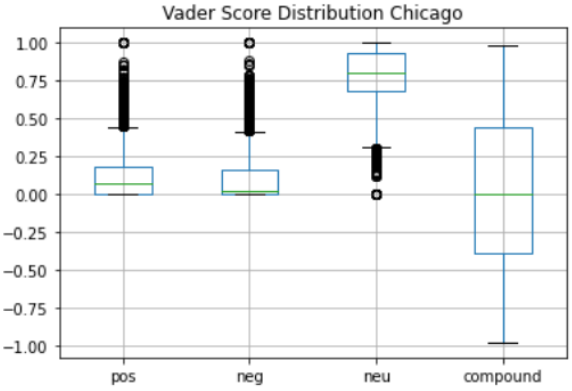
(a) Vader sentiment – NYC



(b) Vader sentiment – LA



(c) Vader sentiment – Houston



(d) Vader sentiment – Chicago

Fig. 1. Box plots of Vader sentiment scores for (a) NYC, (b) LA, (c) Houston, and (d) Chicago.

B. Correlation Analysis

To establish a relationship between sentiment and casualties, a correlation matrix was created that comprised of confirmed cases for each city as shown in Table II, deaths, and the average compound score for each day. This matrix reveals poor correlation between sentiment and casualties. Therefore, there needs to be some grouping of sentiment to refine relationship between major ranges of sentiment and Covid-19 data.

TABLE II: CORRELATION MATRIX OF COVID-19 DATA AND SENTIMENT FOR (A) NYC, (B) LA, (C) HOUSTON, AND (D) CHICAGO

	compound	deaths	cases
compound	1.000000	-0.015946	-0.006857
deaths	-0.015946	1.000000	0.990589
cases	-0.006857	0.990589	1.000000

(a) NYC sentiment vs deaths, cases

	compound	deaths	cases
compound	1.000000	-0.023243	-0.022464
deaths	-0.023243	1.000000	0.996160
cases	-0.022464	0.996160	1.000000

(b) LA sentiment vs deaths, cases

	compound	deaths	cases
compound	1.000000	-0.024541	-0.026321
deaths	-0.024541	1.000000	0.956446
cases	-0.026321	0.956446	1.000000

(c) Houston sentiment vs deaths, cases

	compound	deaths	cases
compound	1.000000	0.082075	0.072874
deaths	0.082075	1.000000	0.997396
cases	0.072874	0.997396	1.000000

(d) Chicago sentiment vs deaths, cases

Following this, the compound sentiment was categorized by the positive, negative, and neutral range, and averaged per day. All data used (Covid-19 data and sentiment data) are normalized between 0 – 100 at this stage. After categorization, there are more definitive correlation values for Covid-19 casualties. Additionally, data points from August 18th to September 2nd were omitted as they contributed to poor correlation. In the following matrices, poscomp is the compound scores that are within the positive range (>= 0.05), negcomp is the compound scores that are within the negative range (<= -0.05), and neucomp is the compound scores that are within the neutral range (< 0.05 and > -0.05). The analysis result for each city is shown in Table III.

TABLE III: CORRELATION MATRIX OF COVID-19 DATA AND RANGE GROUPED SENTIMENT FOR (A) NYC, (B) LA, (C) HOUSTON, AND (D) CHICAGO

	poscomp	negcomp	neucomp	deaths	cases
poscomp	1.000000	0.183753	0.012328	0.041653	0.029490
negcomp	0.183753	1.000000	-0.238506	0.467915	0.493409
neucomp	0.012328	-0.238506	1.000000	-0.005808	-0.001501
deaths	0.041653	0.467915	-0.005808	1.000000	0.988439
cases	0.029490	0.493409	-0.001501	0.988439	1.000000

(a) NYC sentiment grouped by range vs deaths, cases

	poscomp	negcomp	neucomp	deaths	cases
poscomp	1.000000	0.183753	0.012328	-0.038010	0.009072
negcomp	0.183753	1.000000	-0.238506	0.465797	0.489312
neucomp	0.012328	-0.238506	1.000000	0.037098	0.027869
deaths	-0.038010	0.465797	0.037098	1.000000	0.990582
cases	0.009072	0.489312	0.027869	0.990582	1.000000

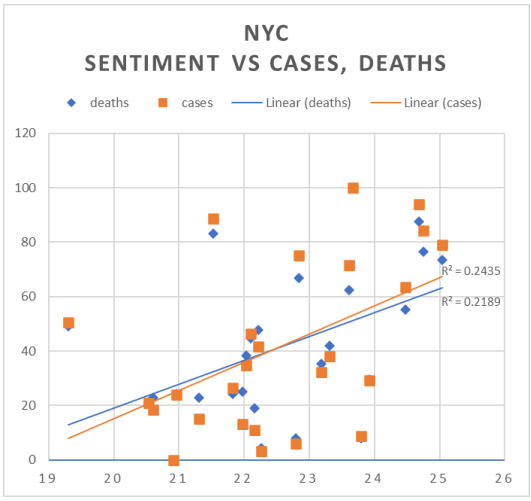
(b) LA sentiment grouped by range vs deaths, cases

	poscomp	negcomp	neucomp	deaths	cases
poscomp	1.000000	0.183753	0.012328	-0.029304	0.004788
negcomp	0.183753	1.000000	-0.238506	0.459895	0.544598
neucomp	0.012328	-0.238506	1.000000	0.072914	-0.010884
deaths	-0.029304	0.459895	0.072914	1.000000	0.964798
cases	0.004788	0.544598	-0.010884	0.964798	1.000000

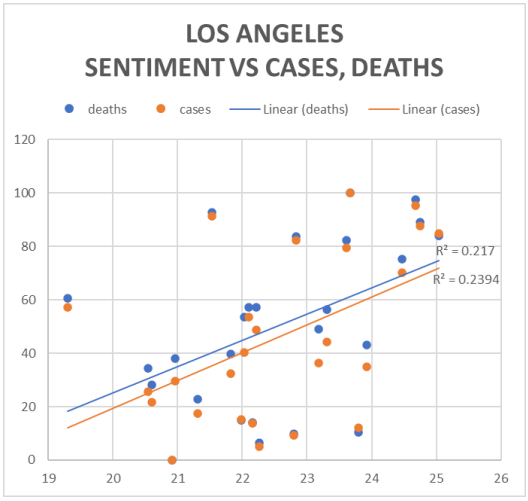
(c) Houston sentiment grouped by range vs deaths, cases

	poscomp	negcomp	neucomp	deaths	cases
poscomp	1.000000	-0.168709	-0.051024	0.249155	0.239678
negcomp	-0.168709	1.000000	-0.086779	0.231964	0.225377
neucomp	-0.051024	-0.086779	1.000000	0.229645	0.219663
deaths	0.249155	0.231964	0.229645	1.000000	0.996860
cases	0.239678	0.225377	0.219663	0.996860	1.000000

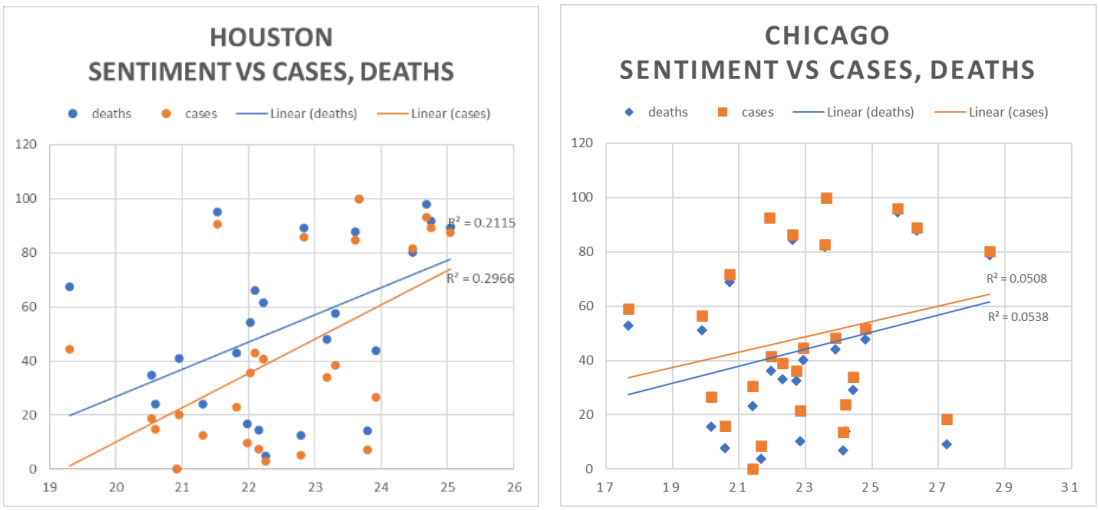
(d) Chicago sentiment grouped by range vs deaths, cases



(a)



(b)



(c) (d)

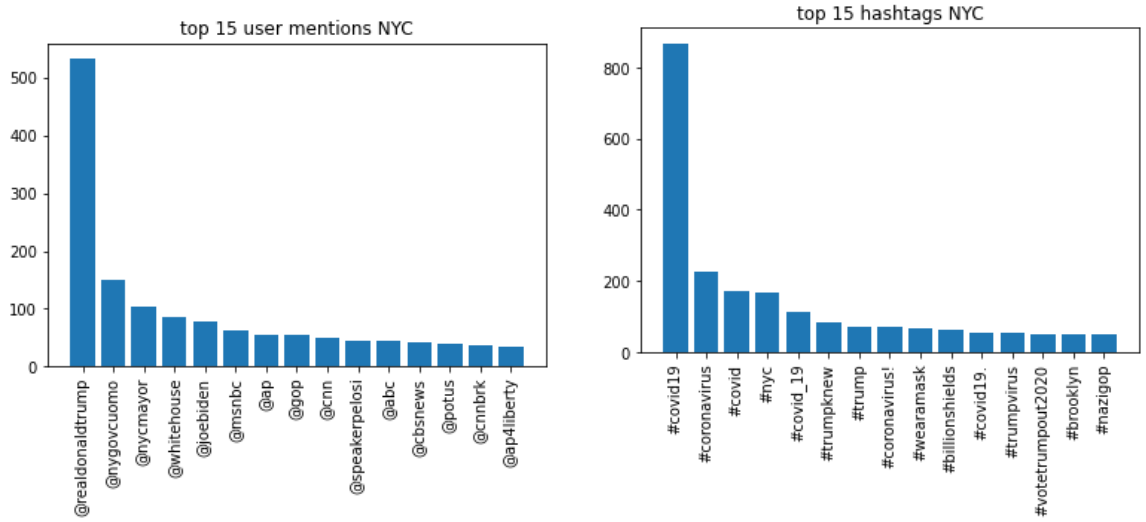
Fig. 2. Plots of relationship between Covid-19 casualties and sentiment for (a) NYC, (b) LA, (c) Houston, and (d) Chicago.

C. Relationship of Negative Sentiment Analysis

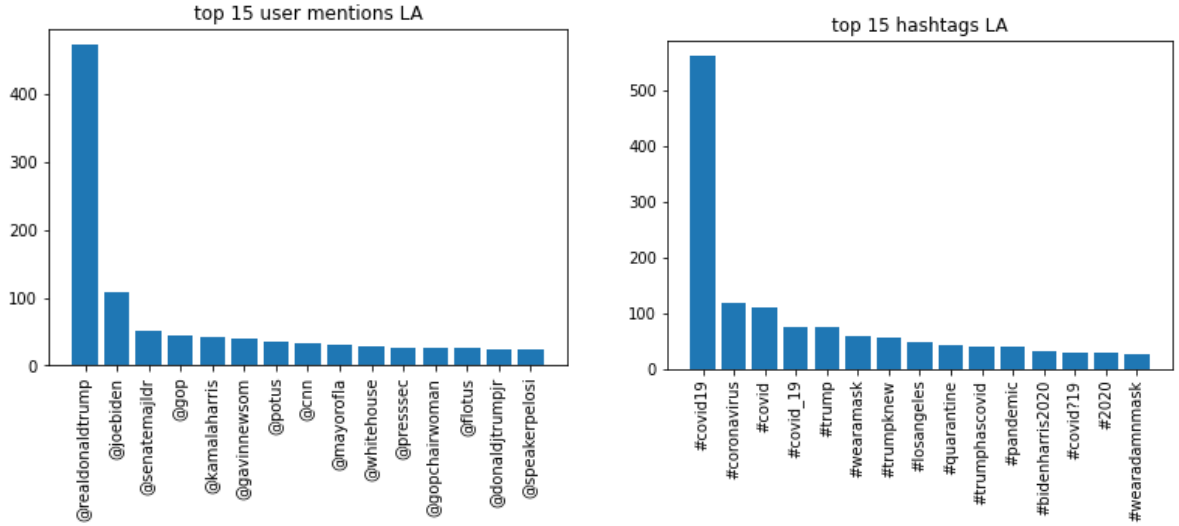
With much stronger relationship between negative sentiment values and Covid-19 casualties, the data for negative sentiment is plotted using Excel as shown in Fig. 2.

D. Analysis of Hashtags and Users

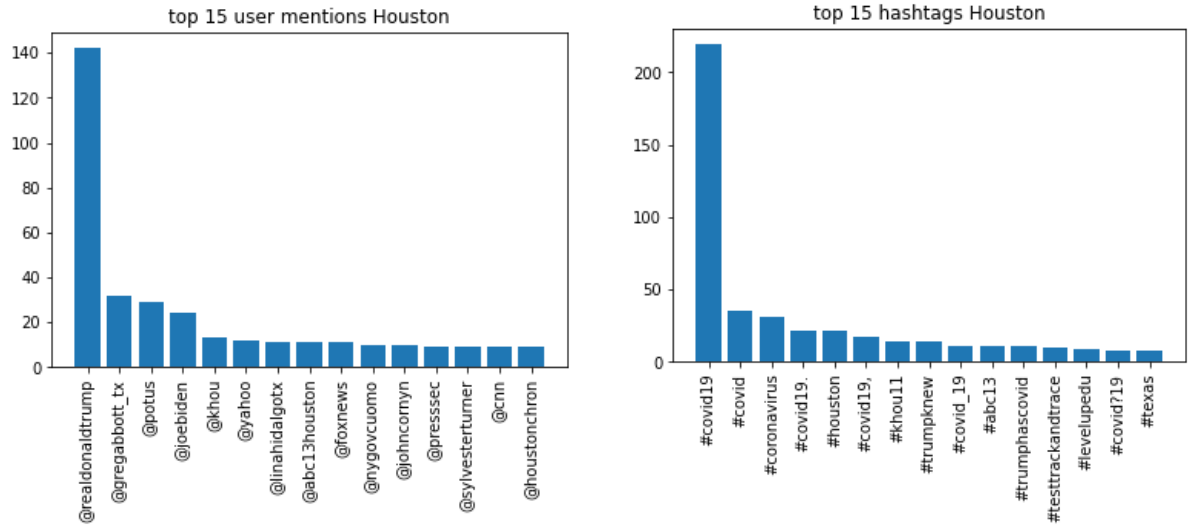
To identify the top 15 most commonly used hashtags and most frequently mentioned users from the tweets, the text is tokenized and only tokens beginning with a '#' or '@' are plotted to show the frequency. All the cities top mentions are of politicians and news outlets with Donald Trump being the most mentioned user. Similarly, top hashtags, apart from #covid19, majorly address President Trump, the city, and the Biden-Harris campaign. The data distribution plots are shown in Fig. 3.



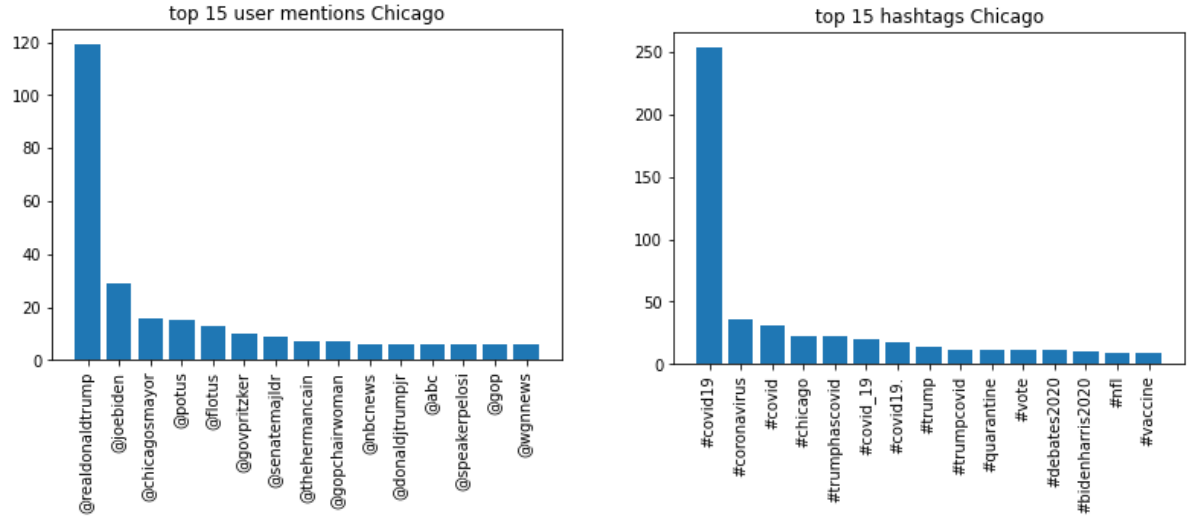
(a) NYC top 15 user mentions and hashtags



(b) LA top 15 user mentions and hashtags



(c) Houston top 15 user mentions and hashtags



(d) Chicago top 15 user mentions and hashtags

Fig. 3. Plots of top 15 most common user mentions and hashtags in (a) NYC, (b) LA, (c) Houston, and (d) Chicago.

E. Classification and Frequency

TABLE IV: THE COUNTS FOR THE CLASSIFIED CATEGORY, WORD LIST FOR (A) NYC, (B) LA, (C) HOUSTON, AND (D) CHICAGO

Category	Words	Count
Covid-19	{covid, covid19, coronavirus, quarantine, corona, pandemic, virus}	7936
Covid-19 related	{positive, mask, deaths, died, test, dead, cases, care, health, tested, masks, vaccine, life, lives, lost}	3716
Political	{trump, president, biden, trumps}	1984
Broader concerns	{people, york, americans, nyc, work, country, school, city, family}	3504

(a) NYC words classified by category

Category	Words	Count
Covid-19	{covid, covid19, quarantine, coronavirus, corona, virus, pandemic}	6107
Covid-19 related	{mask, died, positive, deaths, test, vaccine, masks, die, care, tested, life, cases, testing, lives, dead, health}	2553
Political	{trump, president, trumps}	1209
Broader concerns	{people, americans, work, california, country, angeles, american}	1776

(b) LA words classified by category

Category	Words	Count
Covid-19	{covid, covid19, quarantine, corona, coronavirus, pandemic, virus}	2385
Covid-19 related	{mask, positive, deaths, cases, tested, test, vaccine, died, masks, care, health, life}	931
Political	{trump, president, biden}	395
Broader concerns	{people, houston, texas, work, family, house, school, safe}	825

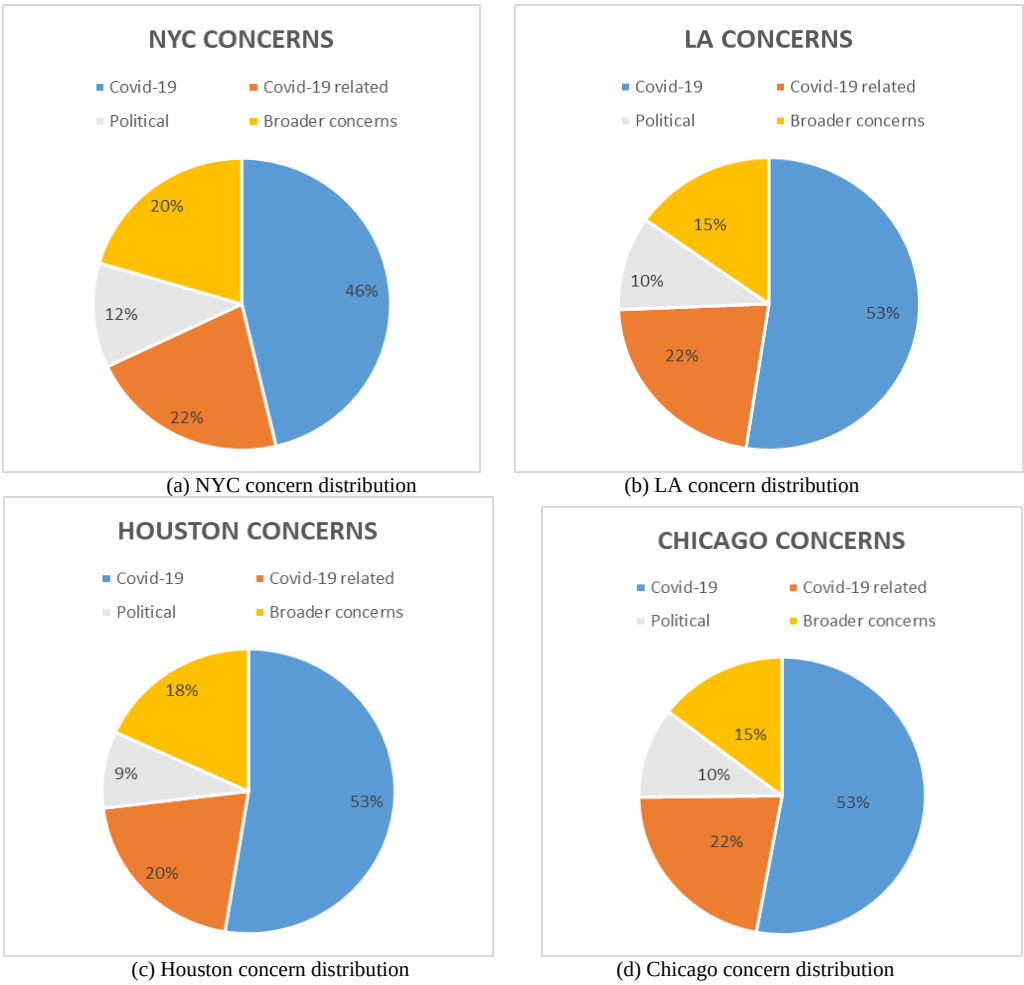
(c) Houston words classified by category

Category	Words	Count
Covid-19	{covid, covid19, quarantine, coronavirus, pandemic, corona, virus}	2339
Covid-19 related	{mask, positive, test, deaths, care, cases, tested, masks, died, health, die, feel, life, vaccine}	971
Political	{trump, president, biden}	468
Broader concerns	{people, chicago, work, americans, country}	645

(d) Chicago words classified by category

To obtain topics that were most frequently tweeted about, the tweet text is tokenized and all stop words [12] are removed from the text. Using Pandas functions to map count to word, the top 50 occurrences are manually analyzed to

obtain topics of concern. As shown in Table IV, most words fall under 4 main categories: Covid-19, Covid-19 related, political, and broader concerns. Words that do not fall under these four major classifications are excluded as they have little meaning out of context.



(a) NYC concern distribution (b) LA concern distribution

(c) Houston concern distribution (d) Chicago concern distribution

Fig. 4. Pie charts of concern category in (a) NYC, (b) LA, (c) Houston, and (d) Chicago.

These categorized words are then plotted as pie charts to better visualize the magnitude of each category of concern.

F. Analysis of Tweets Frequency

Finally, an analysis is done on the frequency of tweets and determine events that could have triggered the number of tweets to have been posted. To accomplish this, the outliers in tweet frequency is obtained using the IQR method. If the frequency for a day is greater than the upper threshold or lower than the lower threshold, trending headlines involving Covid-19 for that day are examined to identify possible causation.

The lower and upper thresholds for all 4 cities are as follows, NYC: (66, 770), Los Angeles: (76.25, 586.25), Houston: (6.5, 262.5), Chicago: (23.25, 249.25) as shown in Fig. 5. The 8 of 10 outliers for all cities fall on the same days: September 24th and October 2nd with Los Angeles and Chicago gaining lower tweet counts on September 20th.

Upon investigation, September 24th saw almost no Covid-19 related headlines, leading to the day being an outlier. Similarly, apart from the headline of the Covid-19 death toll passing 200,000, there was no significant news on September 20th [13]. However, on October 2nd, news came out of President Trump and the First Lady testing positive for Covid-19 and hashtags such as "#TrumpHasCovid" and FLOTUS trending over social media [14].

NYC	date	count	LA	date	count	HOU	date	count	CHI	date	count
18	9/24	18	18	9/24	37	18	9/24	6	18	9/24	9
17	9/20	68	17	9/20	55	17	9/20	37	17	9/20	19
16	9/19	292	19	9/26	182	19	9/26	83	19	9/26	55
2	9/5	302	20	9/27	199	16	9/19	94	21	9/28	76
21	9/28	312	21	9/28	215	2	9/5	100	16	9/19	89
19	9/26	322	3	9/6	239	3	9/6	101	20	9/27	91
3	9/6	330	16	9/19	247	21	9/28	102	3	9/6	96
20	9/27	330	23	9/30	288	20	9/27	103	5	9/8	120
9	9/12	381	24	10/1	294	12	9/15	114	4	9/7	121
12	9/15	383	12	9/15	295	24	10/1	119	12	9/15	124
4	9/7	397	2	9/5	303	22	9/29	125	2	9/5	125
15	9/18	439	22	9/29	303	4	9/7	128	0	9/3	127
23	9/30	442	15	9/18	306	10	9/13	128	15	9/18	129
24	10/1	455	11	9/14	326	15	9/18	131	23	9/30	130
5	9/8	458	4	9/7	328	9	9/12	138	6	9/9	138
6	9/9	468	5	9/8	350	11	9/14	148	24	10/1	141
11	9/14	477	10	9/13	351	6	9/9	153	22	9/29	150
10	9/13	483	9	9/12	368	8	9/11	162	10	9/13	154
22	9/29	503	14	9/17	387	23	9/30	162	1	9/4	159
0	9/3	505	6	9/9	390	1	9/4	163	9	9/12	163
8	9/11	507	13	9/16	400	5	9/8	170	11	9/14	166
1	9/4	516	0	9/3	402	13	9/16	176	14	9/17	166
14	9/17	548	1	9/4	408	14	9/17	179	7	9/10	173
13	9/16	577	8	9/11	443	0	9/3	180	8	9/11	182
7	9/10	598	7	9/10	486	26	10/3	183	26	10/3	195
26	10/3	607	26	10/3	558	7	9/10	213	13	9/16	204
25	10/2	1243	25	10/2	1105	25	10/2	380	25	10/2	381
l=66, u=770			l=76.25, u=586.25			l=6.5, u=262.5			l=23.25, u=249.25		

Fig. 5. Tweet count per day and highlighted outliers.

Trending topics were checked for September 12th, 26th and October 3rd to understand if number of tweets corresponds to trending news. On October 3rd, there is more news of Trump associates testing positive for Covid-19 including his campaign manager, Chris Christie, and Kellyanne Conway [15]. On September 26th, most relevant news is of Florida reopening bars and restaurants with no restrictions in place [16]. And on September 12th, top news is of Trump officials interfering with CDC reports on Covid-19 and Betsy DeVos's rule to reroute Coronavirus aid to private schools getting rejected [17].

IV. Results

By correlating the sentiment data with cases, a relationship is established. There is an insignificant relationship with Covid-19 casualties and positive and neutral compound score averages. However, negative compound score averages had a moderately positive relationship with Covid-19 casualties.

The correlation between positive and neutral compound scores for Chicago is much greater than for other cities, ranging between 0.22 – 0.25, differing from other cities where positive and neutral scores have insignificant values (0.07 or lesser) whereas correlation for negative scores is much lower than that of other cities. This difference may be clarified by deeper analysis of the textual data and close observation of events in Chicago for the time frame.

Upon analyzing tweet text, primary concern is the same among all 4 cities. Covid-19 both as the disease and its impact on social and health aspects tower over the results of this analysis. The categories politics and broader concerns find "people", "americans", "country", and "work" repeated for all cities. As the collection of tweets occurred prior to the 2020 US Presidential Election, the category, broader concerns, can be taken as concerns the population had going forward in the future. This concern towards politics and livelihood is reflected in the analysis of hashtags and mentions, given the trending hashtags can be grouped into politics, Covid-19, and the city.

V. Conclusion

This exploratory analysis uses four ways of studying Tweet data: using correlation to determine relationship, identifying most used hashtags and tagged users, analyzing text to identify frequency of words, and using outlier daily Tweet frequencies to determine possible cause.

In this process, 2 datasets were created for each city: one for tweet sentiment, cases and deaths, and one for tweet text. The sentiment analysis was done on the negative compound sentiments of the tweets, averaged per day and textual examination was done for all the text in the tweet.

The data should be explored more in order to obtain more information about user sentiment during Covid-19. Since the original dataset is extraordinarily large, analyzing the sentiment of all the tweets, localized to city, state, country, or even worldwide, would result in a more reliable conclusion regarding the relationship between sentiment and Covid-19 casualties. Similarly, classifying the tweets by concerns, such as healthcare, politics, economy, etc. can clarify the connection between sentiment and type of concern. Additionally, the retweet data can produce more reliable trackers for sentiment as they contain the favorite count and retweet count.

Conflict of Interest

Both Authors do not have any conflict of interest.

Author Contributions

Amrutha Ragothaman analyzed the raw data, transform the raw data into processed data Using SQL queries, applied BI techniques to create different Visualizations, applied sentiment analysis and Natural language processing Algorithms, used machine learning and data Mining techniques to preform correlation and chi-square test, wrote the paper. Ching-yu Huang guided the overall research and revised the paper. Both authors approved the final draft of the paper.

References

[1] J. Roesslein. Tweepy: Twitter for Python! [Online]. Available: https://Github.Com/Tweepy/Tweepy

[2] Cjhutto. VaderSentiment. [Online]. Available: https://github.com/cjhutto/vaderSentiment

[3] A. Beri. Sentimental Analysis using vader. [Online]. Available: https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[4] M. Morrissey, L. Wasser, J. Diaz, and J. Palomino. (2020). Lesson 3: Analyze word frequency counts using Twitter data and Tweepy in Python. [Online]. Available: https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/calculate-tweet-word-frequencies-in-python/

[5] U.S. county-level data. [Online]. Available: https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv

[6] pandas 1.1.1. [Online]. Available: https://pypi.org/project/pandas/

[7] Zip Code Profile. [Online]. Available: https://web.har.com/zipcode

[8] Metropolitan areas in California. [Online]. Available: https://www.labormarketinfo.edd.ca.gov/definitions/metropolitan-areas.html#list

[9] Greater Chicago regional map. [Online]. Available: https://www.google.com/maps/d/u/0/embed?mid=1zBO2I10kIY305-R2N3AaOmAXsZE&ie=UTF8&hl=en&msa=0&t=m&z=12&output=embed&ll=41.78353658531876%2C-87.61929268354662

[10] Chicago community areas and zip codes. [Online]. Available: https://www.chicago.gov/content/dam/city/sites/covid/reports/2020-04-24/ChicagoCommunityAreaandZipcodeMap.pdf

[11] Python | Sentiment analysis using VADER. [Online]. Available: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/#:~:text=The%20Compound%20score%20is%20a,1%20(most%20extreme%20positive)

[12] sebleier/NLTK's list of English stopwords. [Online]. Available: https://gist.github.com/sebleier/554280

[13] Sunday, September 20, 2020. [Online]. Available: https://www.wincalendar.com/Calendar/Date/September-20-2020

[14] Friday, October 2, 2020. [Online]. Available: https://www.wincalendar.com/Calendar/Date/October-2-2020

[15] Saturday, October 3, 2020. [Online]. Available: https://www.wincalendar.com/Calendar/Date/October-3-2020

[16] Saturday, September 26, 2020. [Online]. Available: https://www.wincalendar.com/Calendar/Date/September-26-2020

[17] Saturday, September 12, 2020. [Online]. Available: https://www.wincalendar.com/Calendar/Date/September-12-2020

Amrutha Ragothaman was a student of School of Computer Science and Technology, Kean University, Union, New Jersey. Her research interests are in the area of data science, data mining, database applications and Python programming. She graduated with a bachelor of science in computer science from Kean University in August 2021 and expected to graduate with a master of science in computer information systems at Kean University in May 2022.

Ching-yu Huang is an assistant professor of the School of Computer Science at Kean University since September 2014. Dr. Huang received a Ph.D. in computer & information science from New Jersey Institute of Technology, Newark, New Jersey, USA.

Prior to joining Kean University, Dr. Huang had more than 16 years of experience in the industry and academics in software development and R&D in bioinformatics. His research focuses SNP genotype calling and cluster detection; image processing and pattern recognition, especially in microarray and fingerprint; geotagged images and location information reconstruction; database application development; data processing automation; e-learning, educational multimedia, methodology, and online tools for secondary schools and colleges. Dr. Huang has more than 30 publications in journals and conferences and more than 40 presentations in workshops and invited lectures.