Zinan Bo. Sentiment Analysis: Relationship between COVID-19 Twitter Sentiment and Daily Confirmed Cases: A Case Study of New York City. A Master's Paper for the M.S. in I.S degree. November 2021. 49 pages. Advisor: Yue Wang

Coronavirus disease 2019 is a new viral disease, named after the year in which it first appeared. On January 30, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic. With the outbreak of the COVID-19 pandemic since the end of 2019, the economy and lives of people around the world have been very severely affected. At the same time, social media has become the main platform for people to express their concerns, opinions and feelings about the pandemic disease. Social media platforms are flooded with a wide variety of information related to COVID-19. As a result, especially during the period of the COVID-19 lockdown, more people were choosing to search for information, express their emotions, and seek peace on social media. Among many social media, Twitter is a popular resource. By analyzing health event data posted on the Twitter platform, researchers can not only get first-hand information about ongoing health events, but they can also get real-time information faster. These can help health professionals and policymakers respond effectively to health-related events.

Therefore, this study performs the sentiment analysis of Tweets posted by people in New York from March to December 2020. 20,980 Tweets are collected, along with a daily dataset of confirmed cases in New York City in 2020. The data was cleaned, organized, and merged via Python, and then calculated the correlation values between the two datasets. There is a negative correlation between people's sentiment and daily confirmed cases during COVID-19. This study aims to analyze public sentiment toward the pandemic to better understand how public emotions and views about the pandemic change over time.

Headings:

Pandemic -- COVID-19

Social Media -- Twitter

Sentiment Analysis

SENTIMENT ANALYSIS: RELATIONSHIP BETWEEN COVID-19 TWITTER SENTIMENT AND DAILY CONFIRMED CASES: A CASE STUDY OF NEW YORK CITY

by
Zinan Bo

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel
Hill in partial fulfillment of the requirements
for the degree of Master of Science
inInformation Science.

Chapel Hill, North Carolina

November 2021

Approved by

Yue Wang

Contents

1.Background and Introduction

1.1 Background

Coronavirus is a novel viral disease that is named to indicate the year in which it first appeared [1]. On January 30, 2020, the World Health Organization (WHO) declared COVID-19 as a pandemic. With the outbreak of the COVID-19 pandemic since the end of 2019, the economy and lives of people around the world have been severely affected. At the same time, social media has become the main platform for people to express their concerns, opinions and feelings about the pandemic disease. Social media platforms are flooded with a wide variety of information related to COVID-19. As a result, especially during the COVID-19 lockdown, more people chose to search for information, express their emotions, and seek peace on social media.

Since January 2020, COVID-19 has been one of the top topics on major social media and continues to be discussed at the time of writing. There has been a tremendous increase in people's reliance on different social media platforms to receive news and express opinions as opposed to traditional news sources and expression methods. The amount of data presented by these social media platforms has led to an increased interest in using information retrieval, sentiment analysis, natural language processing, and artificial intelligence to analyze texts [2]. This information contains different social phenomena such as cultural dynamics, social trends, natural disasters, public health, frequently discussed topics and opinions expressed by people using social media, etc. This information includes different social phenomena such as cultural dynamics, social trends, natural disasters, public health, popular topics and opinions, etc. The comments and experiences shared by end users constitute a rich repository of information, such that public platforms and social media

become prominent sources of information for the study of rapidly evolving public sentiment issues [3]. Therefore, social media create the possibility to analyze the dynamics of public sentiment during the pandemic, revealing insights about prevailing sentiment and network effects.

Among these public social platforms, the use of Twitter for social media research remains highly popular in academia and industry, with no other platform could attract as much attention from research people as Twitter. Although, Twitter is not the most popular platform in terms of monthly active users, ranking eighth in the overall list [4] (see Figure 1). Facebook and WhatsApp have more active users and are top two in the ranking. However, these platforms with high user activity do not provide data on a large scale as the Twitter platform does. No other social media platform has the infrastructure that Twitter has, and this is what makes the Twitter platform unique - its infrastructure allows any user to be able to follow another user, and it provides almost 100 percent of its data through APIs. Therefore, with such a large number of active users and access to raw data, Twitter has become a very popular social media platform for the research industry.

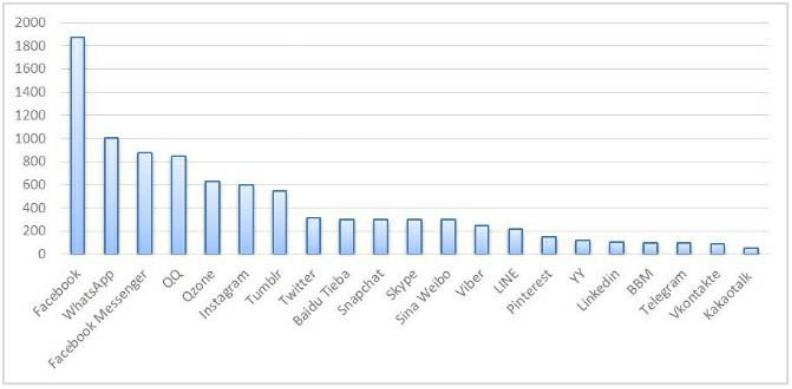


Figure 1: Number (in millions) of monthly active users across social media platforms. Created using data powered by statista.

Twitter data is valuable for revealing public discussion and sentiment related to various topics, as well as real-time news updates during global pandemics, such as H1N1 and Ebola, among others [5-8]. Chew and Eisenach's study [5] shows that Twitter can be used for real-time "information epidemiology" studies as a source of input for health authorities to respond to topics of public concern. During the COVID-19 pandemic, many government officials around the world used Twitter as one of their main communication channels, regularly sharing policy updates and news related to COVID-19 to the public [9]. Therefore, analysis of health event data posted on the Twitter platform not only provides first-hand evidence of the occurrence of health events, but also provides faster access to real-time information to help health professionals and policy makers respond appropriately to health-related events. By analyzing public perception of disease, it is possible to better understand how public sentiment and opinions about disease change over time.

Since there is no vaccine and no cure or approved pharmaceutical intervention for COVID-19 during the year 2020, the fight against the pandemic has been reliant on non-pharmaceutical interventions (NPIs). These NPIs include: (1) case-driven measures such as testing, contact tracing and isolation; (2) personal preventive measures such as hand hygiene, cough etiquette, face mask use, eye protection, physical distancing and surface cleaning, which aim to reduce the risk of transmission during contact with potentially infectious individual; and (3) social-distancing measures to reduce interpersonal contact in the population. In the United States, social-distancing measures have included policies and guidelines to close schools and workplaces, cancel and restrict mass gatherings and group events, restrict travel, maintain physical separation from others (for example, keeping six feet apart) and stay-at-home orders. Non-pharmaceutical interventions and other responses to COVID-19, especially stay-at-home orders, have varied widely across states, leading to spatial and temporal variation in the timing and implementation of mitigation strategies.

This variation in policies and response efforts may have contributed to the observed heterogeneity in COVID-19 morbidity and mortality across states. Also, the scientific evidence whether NPIs are effective hasn't been conclusive during the early 2020. For instance, at the beginning of the pandemic, medical experts lacked good evidence on how SARS-CoV-2 spreads, and they didn't know enough to make strong public-health recommendations about masks. Most of the early evidence was from observational and laboratory studies (indirect evidence) and direct evidence on the efficacy of NPIs has been limited. It has been reported that the general public has been confused due to controversial studies and mixed messages. This has resulted in huge variations in topics and sentiments on NPI measures and the extent to which these measures influence the COVID-19 transmission in different states. Therefore, through the tweets that people post on Twitter, I would like to study the relationship between the sentiment of people and the spread of the epidemic.

1.2 Introduction

Information Retrieval (IR) is a broad field in information science. There is a distinction between information retrieval in a broad sense and in a narrow sense. Information retrieval in the broad sense is called "information storage and retrieval", which refers to the process of organizing and storing information in a certain way and finding the relevant information according to the needs of users. Information retrieval in a narrow sense is usually called "information search", which refers to the process of finding the relevant information needed by the user from the information collection. Whether in the broad or narrow sense of information retrieval, it is an important part of information retrieval to query the corresponding system for documents matching its corpus according to the user's information needs, and to return relevant documents using some selected models. Many modern IR systems choose to use natural language for querying, because such querying is user-friendly and can better popularize IR systems. Therefore, text processing is an

important part of IR. Sentiment analysis is an important part of the natural language processing field. It intersects with computational linguistics and is used to extend various aspects of information retrieval. This field has become a very active area of research and will continue to grow rapidly.

Sentiment analysis, also known as opinion mining, is a general term for a set of technical concepts aimed at analyzing the sentiment/valence, emotion, evaluation, and attitude that humans have toward a target object [10]. Target objects include, but are not limited to, goods, services, organizations, individuals, events, etc. Sentiment analysis may represent different technical applications in different scenarios, such as sentiment recognition, sentiment classification, opinion mining, opinion analysis, opinion extraction, subjective analysis, sentiment computation, evaluation analysis, etc. In summary, sentiment analysis is the analysis of the human viewpoint embedded in a target object. It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining, and text mining. [11]

Opinions are at the heart of almost all human activity and are a key influence on human behavior. People's beliefs and perceptions of reality, as well as the choices they make, are to a considerable extent influenced by how other people view and evaluate the world. For this reason, people tend to seek the opinions of others when they need to make decisions. Therefore, an important part of sentiment analysis has always been to understand what other people think, and that is the reason that the study of sentiment analysis has spread from the field of computer science to management science and social science and is being applied to almost all areas of business and society.

In the business world, sentiment analysis is often applied in obtaining and analyzing product reviews. This approach not only helps businesses to better understand their users and improve their products, but also helps users to obtain more information about their products. According to a

survey of more than 2,000 American adults, the Internet can help people gather the opinions and experiences of large groups of people who are neither acquaintances nor well-known professional critics [12].

- Eighty-one percent of Internet users (or 60 percent of all Americans) have done online research on a product at least once.
- Twenty percent (15% of all Americans) do so on a typical basis.
- Between 73% and 87% of readers of online reviews of restaurants, hotels, and various services (such as travel agents or doctors) s reviews have had a significant impact on their purchases.
- Consumers report they are willing to pay 20% to 99% more for a five-star rated item than a four-star rated item (the difference stems from the type of good or service being considered).
- Thirty-two percent have rated a product, service or person through an online rating system, and 30 percent (including 18 percent of online seniors) have posted a review or opinion about a product or service online.

This indicates that more and more people are willing to give their opinions to strangers via the Internet. Therefore, industrial activity around sentiment analysis is also booming. Numerous start-ups have emerged, and many large companies have built their own internal sentiment analysis systems. But the consumption of products and services is not the only motivation for people to seek or express their opinions online. The expression and demand for political and current events in society is also another important purpose for people. For example, Tumasjan et al. (2010) applied sentiment analysis to analyze people's opinions about politics on social media. They studied 100,000 Tweets about a political party during the German federal elections and found that

there was a strong relationship between the sentiment expressed on Twitter and the political positions mentioned in the tweets [13]. Therefore, the authors argue that Twitter may reflect the current political landscape and can be used as a valid real-time guide about political sentiment (similar paper). Thus, these surveys and studies suggest that the use of sentiment analysis in a large number of domains can be very helpful for real-world applications.

The use of sentiment analysis can be applied to many fields, including management science, political science, economics, and social science. But there are still many challenges that need to be studied and solved. For example, in mining opinions, researchers need to clarify opinion classification and how to label them. In context mining, besides some obvious words that indicate emotions ("awesome", "terrible", "happy", "sad", etc.), there are also words that need to be put in context to understand their meaning. In context mining, besides some obvious words that indicate emotions ("awesome", "terrible", "happy", "sad", etc.), there are also words that need to be put in context to understand their meaning. For example, "The battery life of this camera is very short", "The focus time of this camera is very short". The same vocabulary, "very short", can take on different meanings in different contexts. An increasing number of opinion mining methods have emerged for different types of content. Thus, the use of opinion mining to study different types of information, such as social media and health-related information, is growing rapidly in volume.

Although linguistics and natural language processing (NLP) have a long history, it is only since 2000 that the field of sentiment analysis has become a very active area of research [11]. One of the important reasons for this is the development of social media. For the first time in human history, we have a huge amount of opinion data recorded in digital form in social media on the Web for analysis. Without this data, much research would not be possible. As a result, sentiment analysis is now at the center of social media research.

The purpose of this study was to study the relationship between the sentiment scores of COVID-19-related tweets posted by people in New York City and the daily number of confirmed cases of COVID-19 in New York City between March and December 2020.

Therefore, this paper seeks to answer the following research question:

> Is there a correlation between the sentiment expressed by people in New York City on Twitter and the spread of COVID-19 in New York City?

If the answer is “yes”, then it implies that social media conversation and public health trend are correlated. This may open up the opportunity for further exploration, such as forecasting public health trends and analyzing causal mechanisms. If the answer is “no”, then it is an equally informative negative result. Therefore, it is worth studying this research question.

2. Literature Review

2.1 Sentiment Analysis

Bo Pang and Lillian Lee are two of the leading researchers in sentiment analysis. Lee is currently a professor in computer science at Cornell University. Her main research areas are Natural Language Processing (NLP) and social interactions (Lillian Lee: Research Summary). Pang is currently working at Google, and he focuses on Natural Language Processing (NLP) and social media (Bo. pang). "Opinion Mining and Sentiment Analysis," [15] is their monograph that has been cited throughout the field. It is a detailed monograph covering the background of sentiment analysis, relevant examples of application, technical challenges, and approaches. The challenges they discuss regarding sentiment analysis include, subtlety of sentiment, the differences between fact finding Information Retrieval and opinion mining, and domain context. The approaches they discuss to sentiment analysis include unsupervised methods, domain adaptation, and relationship classification. Lastly, they discuss that the implications for broader consideration in sentiment analysis are privacy and manipulation.

Sentiment analysis is also referred to as opinion mining, while many researchers have subtle differences in the definitions of "sentiment" and "opinion". But Pang and Lee (2008) conclude that "sentiment" or "opinion" is most often defined as subjective opinions that cannot be verified or objectively observed [15]. They define "polarity" to express the subjective text of a positive or negative opinion. They define " strength" to express how strongly the opinion is expressed. For example, if a book is evaluated as "This book is great!", then such a rating has a stronger sentiment than a rating such as "The book is great.", because the word "great" is more positive than the word

"good", and the exclamation mark indicates more excitement. Therefore, both polarity and intensity need to be taken into account when analyzing sentiments.

Pang and Lee (2008) consider that extracting sentiment from text is very different from fact-based textual analysis. Fact-based textual analysis determines the topic of a document by using term frequency, tf*idf, etc., and then categorizes the document according to different topics, while sentiment classification requires different approaches. For example, in sentiment classification, we usually generalize a relatively few classes to many different domains and users. As Pang and Lee say, "… with sentiment classification, we often have relatively few classes (e.g., "positive" or "3 stars ") that generalize across many domains and users.... ...In fact, the regression-like nature of strength of feeling, degree of positivity, and so on seems rather unique to sentiment categorization " [15]. So, if we use binary classification, the labels of data are opposing (e.g., positive/negative). If we use ordinal categories, sentiments will be expressed in a small range (e.g., we can use a five-star rating system to rate all kinds of different products or services on different websites). These methods of sentiment classification do not change as the subject of the document changes. Pang and Lee (2008) note that the development of labeled data has brought large-scale empirical evaluation tools to the field of sentiment analysis, essentially having the "right" analysis to test new systems [15].

Domain context is another challenge with Sentiment analysis. One reason for this issue is that relying on keywords to determine sentiment can be a challenge. This is because the positivity or negativity of some words can imply different meanings in different domains. Some words may have positive meanings for one domain but may have negative meanings for other domains. As Pang and Lee (2008) put it, “Compared to topic, sentiment can often be expressed in a more subtle manner, making it difficult to be identified by any of a sentence or document’s terms when

considered in isolation" [15]. For example, if a concert is described as "crazy", the sentiment expressed behind this might be positive for fans. But if a pandemic is described as "crazy," then the meaning behind it is negative.

- The domain context is different from the textual context. The textual context is based on the text to understand the phrases, but in domain context, it is the domain in which the sentiment is expressed that is key to understanding. For example, in a book review, if the review of a book is "go read the book", then the review is positive. But if a movie review says, "go read the book", then the evaluation probably is negative [15, p.13]. As Pang and Lee (2008) put it, "In general, sentiment and subjectivity are quite context-sensitive, and, at a coarser granularity, quite domain dependent (in spite of the fact that the general notion of positive and negative opinions is fairly consistent across different domains)" [15, p.13]. Therefore, although lexicons can be applied in different domains, if a word has different sentiments in different domains, then lexicons cannot be used as the only source of information about sentiment.
- In addition, gathering, organizing, and maintaining the list of words in a lexicon takes a long time. For example, determining which words should be included, labeling which words in the word list as positive or negative, etc. can be problematic. This requires research to determine which words have strong enough sentiment across domains to be included in the lexicon, and that list of terms must also be maintained with care to exclude redundant terms. If these were to be collected and maintained manually, it would be an extremely labor-intensive and resource-intensive task. Machine learning is one good way to alleviate this problem by training the data to achieve higher accuracy in analysis [15, p.11].

There are two unsupervised machine learning approaches to sentiment analysis, lexicon-based and bootstrapping. Lexicon-based unsupervised learning approach creates a sentiment lexicon in an unsupervised manner and then determines the degree of positivity or subjectivity of a text unit by some function based on the positive and negative indicators identified by the lexicon." [15] There are also some variants of this approach. For example, words could be collected based on whether they appear with other words (using mutual information and co-occurrence), and seed words are used to determine which clusters to label as positive or negative. Bootstrapping uses the results from the initial classifier to create training data and applies a second algorithm to the result. Each of these algorithms can help systems train data by themselves and then provide sentiment analysis.

- One of the most popular lexicon-based sentiment analysis methods is Linguistic Inquiry and Word Count (LIWC)。Linguistic Inquiry and Word Count (LIWC; pronounced "Luke") is a text analysis program that calculates the percentage of words in a given text that fall into one or more of over 80 linguistic, psychological and topical categories indicating various social, cognitive, and affective processes. We can use LIWC, for example, to determine the degree in which a text uses positive or negative emotions, self-references or causal words. The core of the program is a dictionary containing words that belong to these categories. Dictionaries for many languages are available; it is also possible to define your own dictionary, for example to define one or more categories that are not included in the standard dictionary.
- Another lexicon-based tool is SentiWordNet, a lexical resource designed to support sentiment classification and opinion mining applications, which has evolved into its third version. SentiWordNet 1.0 is publicly available for various research projects worldwide and is currently licensed to more than 300 research groups [16]. It automatically annotates

all synonyms in WordNet according to the degree of positivity, negativity and neutrality of the words. For example, blasphemous, blue, and profane are all in the same synset because they meet the definition of "characterized by profanity" [17]. SentiWordNet addresses three main problems in sentiment analysis: determining sentiment, determining objectivity, and determining polarity intensity. Thus, each word in a phrase is assigned three scores: one for positive sentiment, one for negative sentiment, and one for objectivity. These scores range from 0.0 to 1.0 and add up to 1.0 [17]. SentiWordNet 3.0, in contrast to SentiWordNet 1.0, the algorithm used to automatically annotate WordNet now includes, in addition to the previous semi-supervised learning step, a random-walk for refining the scores step. The result of the study shows an improvement in accuracy of about 20% compared to SentiWordNet 1.0 [16].

- VADER (Valence Aware Dictionary and Sentiment Reasoner) also is a lexical and rule-based sentiment analysis tool. Because it is sensitive to both the polarity (positive/negative) and the strength of sentiment, it is well suited for analyzing sentiment expressed in social media. It is available in the NLTK package and also can be applied directly to unlabeled text data. VADER sentiment analysis relies on a comprehensive lexicon of sentiments that are usually labeled as positive or negative based on their semantics. It could tell us not only the positivity and negativity scores, but also the degree of positivity or negativity of an emotion. The sentiment score of a text can be obtained by summing up the strength of each word in the text. For example, words like "love", "enjoy", "happy", and "like" all express a positive emotion. At the same time, VADER is smart enough to understand the basic context underlying these words, such as "do not love" which is a negative statement. And

it also understands the importance of capitalization and punctuation. Therefore, VANDER is a suitable tool for studying social media sentiment analysis.

Relationship classification is the last approach described by Pang and Lee. There are many relationships to consider when classifying sentiment, and the relationship between users and user communication is one of the relationships to be taken into account. Pang and Lee found that in a study of 100 responses in a newsgroup, a discourse relationship consisting of opposing sentiments emerged. For example, if one user responded negatively to an article, then another user's response to that user could be a positive response about that article, and a response to that positive response could be a negative response [15]. Therefore, understanding such trends can help develop tools more effectively when building analysis tools. Also, the relationship between sentences and documents is important to be considered. For example, "I really like the food in this restaurant, but I don't like the service in this restaurant" shows two opposing sentiments in a sentence. Therefore, by monitoring these different emotions, it is possible to assign objectivity to sentences in the document, or to monitor the emotions of the whole document [15].

2.2 Social Media

Social media is a virtual community and online platform for people to create, share, and exchange opinions, ideas, and experiences. Social media gives users more choice and editing power and allows them to assemble themselves into a kind of reading and listening community. Social media can also be presented in many different forms, including text, images, music and video. So social media is an emotionally rich field. More and more companies are choosing to reach out to their users on social media, to do user data research and user support, and to allow users to participate in the development of their products or services. And around the world, more and more government departments and officials are using social media as one of their main communication

channels, sharing policy updates and news to the public on a regular basis. Therefore, based on the large number of emotions expressed on social media, using social media data for sentiment analysis can help to understand people's emotions and attitudes towards social events, which can be very useful for understanding public sentiment.

Although social media is a good platform for sentiment analysis and we already have many tools and methods for sentiment analysis, there are still many challenges in conducting sentiment analysis on social media [18]. Maynard's article mentions Relevance, Target identification, Negation, Contextual information, Volatility over Time, Opinion Aggregation and Summarisation, as challenges related to sentiment analysis on social media.

- Relevance

  "Even when a crawler is restricted to specific topics and correctly identifies relevant pages...discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand" [18, p.18]. People's expressions on social media are more autonomous and diffuse, and there are off-topic discussions even on pages of related topics, which makes sentiment analysis difficult. This can be solved by trying to train a classifier for relevant topics or comments, for example, by removing comments that contain certain terms that are not needed. Furthermore, clustering can be used to find sentences or segments with opinions related to certain topics and ignore those that do not belong to those topics. However, these two methods may miss some relevant comments.

- Target identification

  "One problem faced by many search-based approaches to sentiment analysis is that the topic of the retrieved document is not necessarily the object of the sentiment held therein"

[18, p. 24]. This means that there is probably no connection between the keywords searched and the opinions expressed by the user. For example, the day after Whitney Houston's death, TwitterSentiment and some similar sites showed that the majority of tweets about Whitney Houston were negative. But these negative views were expressing sadness about the event, not expressing that they did not like Whitney Houston. therefore, instead of just trying to decide what the sentiment was without reference to the target, one could try to first identify the relevant topic (target/entity) that expresses the sentiment, and then look for semantically related views to that entity to solve the problem. Mark documents as containing sentiment (instead of marking them as including the topic of sentiment). This eliminates the need to group them into a topic and still be able to retrieve results that include sentiment.

- Negation

  Some simpler word-package sentiment classifiers do not handle negation well. the Unigram-based approach would make sentiment judgments by judging one word at a time, which would cause the difference between the phrases "bad" and "good" to be ignored. One solution is to add more features, such as n-grams or depending on structures. Another solution is to capture simple patterns by inserting single-piece words like "NOT-helpful" and "NOT-exciting", avoiding the need for analysis [18, p. 24-25].

- Contextual information

  “Social media, and in particular tweets, typically assume a much higher level of contextual and world knowledge by the reader than more formal texts” [18, p.25]. Thus, this raises difficulties when contextual information needs to be collected in social media to fully understand some comments. For example, Maynard et al give an example in their article

where a user compares a politician to a fictional character in a novel. However, such a comment might not be easily understood by automatic methods. Therefore, this problem can be mitigated by considering the use of metadata on social media. For example, Twitter has a large number of metadata related to tweets posted by users. This metadata can help in aggregating and summarizing users' views, and it can also help in removing ambiguities and training data.

- Volatility over Time

  In social media, especially Twitter, the opinions expressed by users can change radically over time, from positive to negative, and vice versa. Therefore, metadata may be useful when it comes to the situation in social media where users' opinions fluctuate over time. The use of timestamps, as pointed out by Maynard et al. (2012), is one way to address this issue, a method that places sentiment in the correct temporal context. For example, since the beginning of 2020, when the COVID-19 Pandemic outbreak, people's sentiment has changed significantly over time. The emotions that users show on social media in 2020 are not representative of that their emotions toward COVID-19 in 2021. Therefore, Maynard et al. suggest that the opinions and sentiments of users extracted from social platforms could be timestamped and then stored in a knowledge base. This knowledge base is continuously enriched as new content and opinions emerge. However, it is a challenge to detect newly emerged opinions. And the contradictions and changes that people show on social media over time need to be captured.

- Opinion Aggregation and Summarisation

  “Opinions behave differently here, however: multiple opinions can be attached to an entity and need to be modelled separately, for which we advocate populating a knowledge base”

[18, p.25]. Maynard et al. consider one of the important questions to be whether it is appropriate to only store the average of the opinions detected over a given time interval, or to store more detailed information, "such as modeling the source and intensity of conflicting opinions and how they change over time [18, p.25]." In their article, they advocate storing an opinion-based summary, such as a timeline that shows positive/negative opinions with opinion holders and key characteristics. And it is possible to cluster the opinions expressed by users in social media through information about demographics, etc. Thus, the nature of social media (interactive, graph-based, etc.) requires new approaches to opinion aggregation.

3.Methodology

I chose to study the relationships between tweets posted by people in New York City, and daily confirmed cases of COVID-19. There are many advantages to study data from such a metropolitan area, for example, such a big city has sufficient amount of data and information to conduct the research. And the dataset available is also very informative. I first downloaded data on English-language tweets related to COVID-19 posted by people in New York City on Twitter from March 2020 to December 2020, and then downloaded the daily records of confirmed cases provided by New York City on its official website. After understanding the two datasets, I cleaned, organized and merged the data by using python. Finally, the relationship between these data was calculated by correlation.

3.1 Twitter Data

In the early period of the COVID-19 pandemic outbreak, it has been difficult to use social media resources like Twitter to study related issues. Because some of the Twitter datasets on COVID-19 released at the time included a wide range of topics and domains [19-22], such datasets were not user-friendly for researchers to utilize. Researchers would also need to understand and clean the data before using them, which would involve a lot of additional time and effort. Sara Melotte and Mayank Kejriwal aim to use the Twitter datasets they created for 10 metropolitan cities to help researchers be able to study the COVID-19 epidemic in a metropolitan context through the lens of social media [23].

GeoCOV19Tweets is a dataset of English-language tweets spanning the globe collected by monitoring more than 90 keywords and hashtags frequently used in reference to the COVID-19

pandemic [21]. The data in this dataset was obtained by filtering English tweets from the Twitter streaming API. The collection started in March 2020, with each collection beginning between 10:0 and 11:00 h GMT+5:45 each day [24] and updated daily with newly collected tweet IDs.

The datasets created by Sara Melotte and Mayank Kejriwal are sub-datasets of GeoCOV19Tweets, each containing information on the date, hashtag, and city, state, and type of place where the tweet originated, and also retaining the sentiment scores contained in the GeoCOV19Tweets dataset.

3.1.1 Hydrating tweet IDs

Obtaining tweets from Twitter is not difficult. Researchers can access Twitter's live feeds (streaming API) or TweetSets, which are dehydrated tweets, through Twitter's application programming interface (API) or third-party databases. This means that instead of receiving a file containing the tweets, location, date, image, and other additional information about the tweets, researchers initially receive a file consisting of a list of unique tweet IDs. This is because, although Twitter allows researchers to access and extract data from real-time feeds or search and extract older tweets, Twitter's developer policy do not allow the raw data to be shared with third parties (Twitter’s developer policy, which can be helpful to learn more information). Therefore, only the Twitter ID, user ID and/or message ID can be shared publicly.

The process of retrieving the complete tweets by their IDs is known as hydrating of tweet IDs. The large size of the data set consisting of a large number of related Tweets might another reason why Twitter only provides dehydrated IDs. In this way, a file containing only a series of IDs (numbers) is easier to manage than a csv file containing thousands of tweets and their metadata.

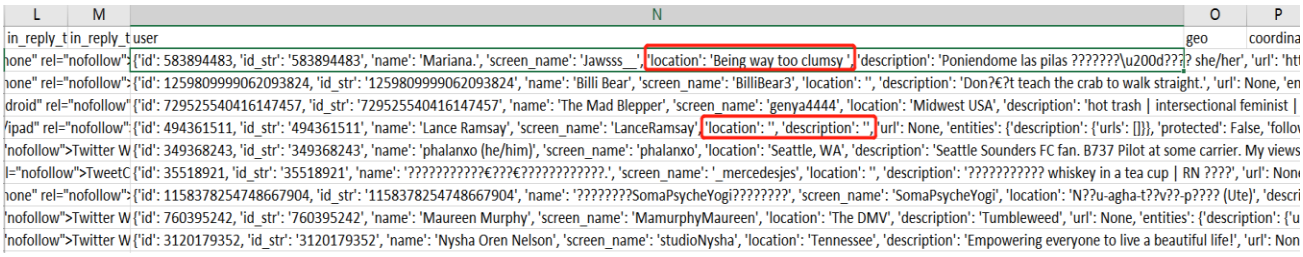
Before accessing Twitter's API, we need to go to Twitter's developer portal and sign up for a developer account. After the account is approved, users can then use third-party tools to access

Twitter's API and hydrate the ID. There are many third-party tools to access the Twitter API, such as the twarc python library, or Hydrator Desktop Application, or DocNow Hydrator Desktop Application, etc. There is also a limit to the number of Tweets that can be retrieved by ID on the Twitter API in a day. The calculation is like this:

> the total (number of 15 min windows in a day) * the number of requests allowed per window * max number of tweets that can be retrieved in every request

Users are not allowed to increase the number of Tweets they can download in a day unless they pay to improve their user service.

However, there are some difficulties in organizing the Tweets data retrieved from the Twitter API. For example, at the beginning of the study, I downloaded ID files from the IEEE website of users across the U.S. in 2020 regarding COVID-19 tweets, as shown in the figure below. I then used these IDs to retrieve the corresponding Tweets from the Twitter API. However, while processing the data, I ran into a problem that many Twitter users would fill in their moods, or symbols, or non-real information on the location profile. As a result, it is difficult to sort out the COVID-19-related Tweets posted by people in New York City from this large amount of data. Therefore, I replaced this dataset with another dataset of tweets with valid location information identified by two researchers using a Reverse-Geocoding tool. Figure 2 shows that the user filled in the location profile with unreal information, or blank information.

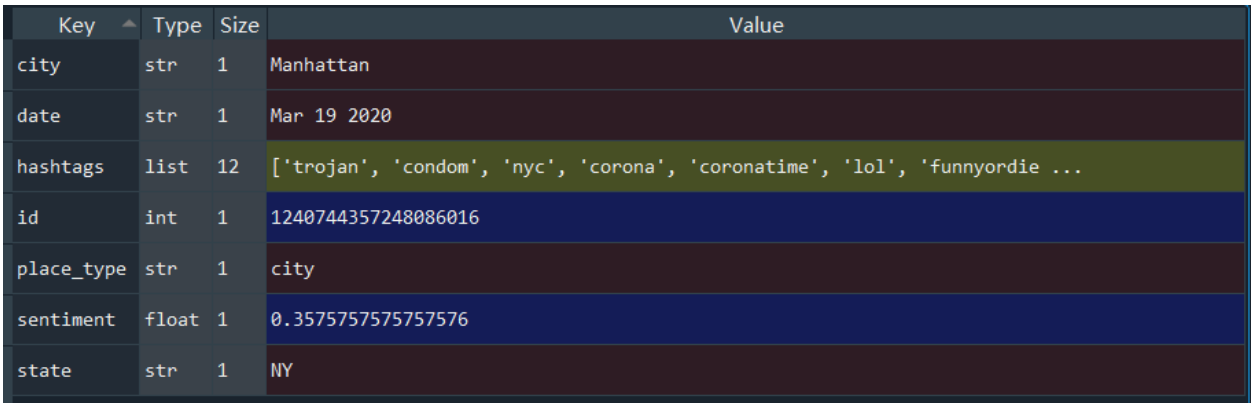


L	M	N	O	P
in_reply_t	in_reply_t	user	geo	coordina
hone" rel="nofollow">		{'id': 583894483, 'id_str': '583894483', 'name': 'Mariana.', 'screen_name': 'Jawsss__', 'location': 'Being way too clumsy ', 'description': 'Poniendome las pilas ???????\u200d???	? she/her', 'url': 'htt	
hone" rel="nofollow">		{'id': 1259809999062093824, 'id_str': '1259809999062093824', 'name': 'Billi Bear', 'screen_name': 'BilliBear3', 'location': '', 'description': 'Don?€?t teach the crab to walk straight.', 'url': None, 'en		
droid" rel="nofollow"		{'id': 729525540416147457, 'id_str': '729525540416147457', 'name': 'The Mad Blepper', 'screen_name': 'genya4444', 'location': 'Midwest USA', 'description': 'hot trash | intersectional feminist |		
/ipad" rel="nofollow"		{'id': 494361511, 'id_str': '494361511', 'name': 'Lance Ramsay', 'screen_name': 'LanceRamsay', 'location': '', 'description': '', 'url': None, 'entities': {'description': {'urls': []}}, 'protected': False, 'follov		
'nofollow">Twitter W		{'id': 349368243, 'id_str': '349368243', 'name': 'phalanxo (he/him)', 'screen_name': 'phalanxo', 'location': 'Seattle, WA', 'description': 'Seattle Sounders FC fan. B737 Pilot at some carrier. My views		
l="nofollow">TweetC		{'id': 35518921, 'id_str': '35518921', 'name': '???????????€???€????????????.', 'screen_name': '_mercedesjes', 'location': '', 'description': '??????????? whiskey in a tea cup | RN ????', 'url': Non		
hone" rel="nofollow">		{'id': 1158378254748667904, 'id_str': '1158378254748667904', 'name': '????????SomaPsycheYogi????????', 'screen_name': 'SomaPsycheYogi', 'location': 'N??u-agha-t??v??-p???? (Ute)', 'descri		
'nofollow">Twitter W		{'id': 760395242, 'id_str': '760395242', 'name': 'Maureen Murphy', 'screen_name': 'MamurphyMaureen', 'location': 'The DMV', 'description': 'Tumbleweed', 'url': None, 'entities': {'description': {'u		
'nofollow">Twitter W		{'id': 3120179352, 'id_str': '3120179352', 'name': 'Nysha Oren Nelson', 'screen_name': 'studioNysha', 'location': 'Tennessee', 'description': 'Empowering everyone to live a beautiful life!', 'url': Non		

Figure 2. An example of a user filling in a location profile with unreal information, or blank information.

3.1.2 Dataset

The dataset is a Java Script Object Notation (JSON) file. The whole of the file is a list, and each element of the list is a dictionary. Each dictionary represents a tweet, and each tweet records: tweet ID, sentiment rating, date, hashtag, city, state, and location type.



Key	Type	Size	Value
city	str	1	Manhattan
date	str	1	Mar 19 2020
hashtags	list	12	['trojan', 'condom', 'nyc', 'corona', 'coronatime', 'lol', 'funnyordie ...
id	int	1	1240744357248086016
place_type	str	1	city
sentiment	float	1	0.3575757575757576
state	str	1	NY

Figure 3. An example of a dictionary in a JSON file

The workflow of this dataset for the Twitter data collection method is shown in the figure 4.

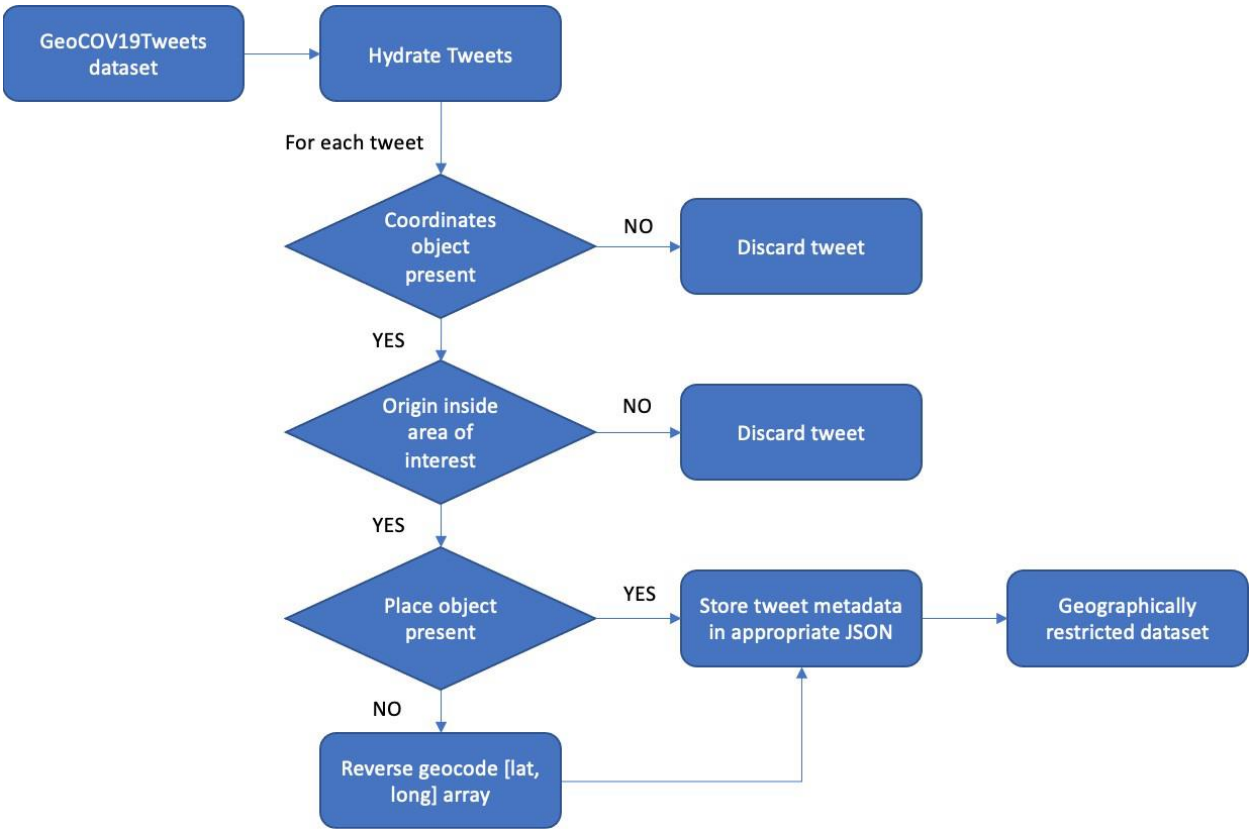


Figure 4. The workflow of this dataset

- Hydrating Tweets

  They first used the Python twarc library [19] to hydrate the tweet IDs in GeoCOV19Tweets from March 20, 2020, to December 1, 2020, 255 days, and then filtered the Twitter data based on the location profile.

- Determining Tweet Origin

  They are interested in the location of the tweet, rather than the user-defined location tag, because the two types of location information are different in many cases. So, after they finished hydrating Twitter IDs, they kept tweets with "coordinate" objects in the metadata and filtered out tweets that did not have "coordinate" objects defined in the metadata.

- Reverse-Geocoding

Even though some tweets have "coordinate" objects defined in the metadata, their "location" may still be blank. So, in this case, the two researchers who created this database used the Geocodio tool to reverse geocode the longitude and latitude in the "coordinates" object.

Geocodio was created by Michele and Mathias Hansen, a married couple from Arlington, Virginia. Geocodio's API allows forward and reverse geocoding within the United States and Canada, returning up to five possible matches with accuracy rankings between 0.00 and 1.00. When the tweet metadata contains only "coordinate" objects and not "location" objects, Sara Melotte and Mayank Kejriwal use the highest precision reverse geocoding results to infer the city, state, zip code, and country of the location.

- Location-filtering

  Sara Melotte and Mayank Kejriwal collected tweets about COVID-19 from people in 10 cities. These cities were the 10 most populous cities in the United States and Canada, namely New York, Los Angeles, Toronto, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, and Dallas. This includes the data I wanted to study, which is the tweets of people in New York City.

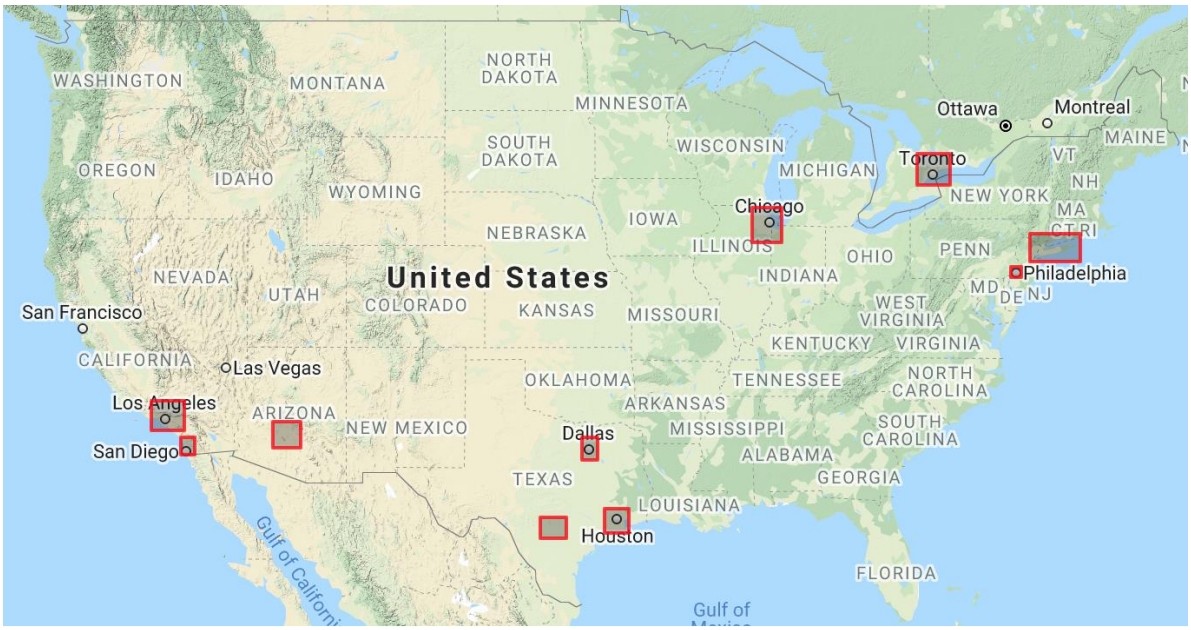


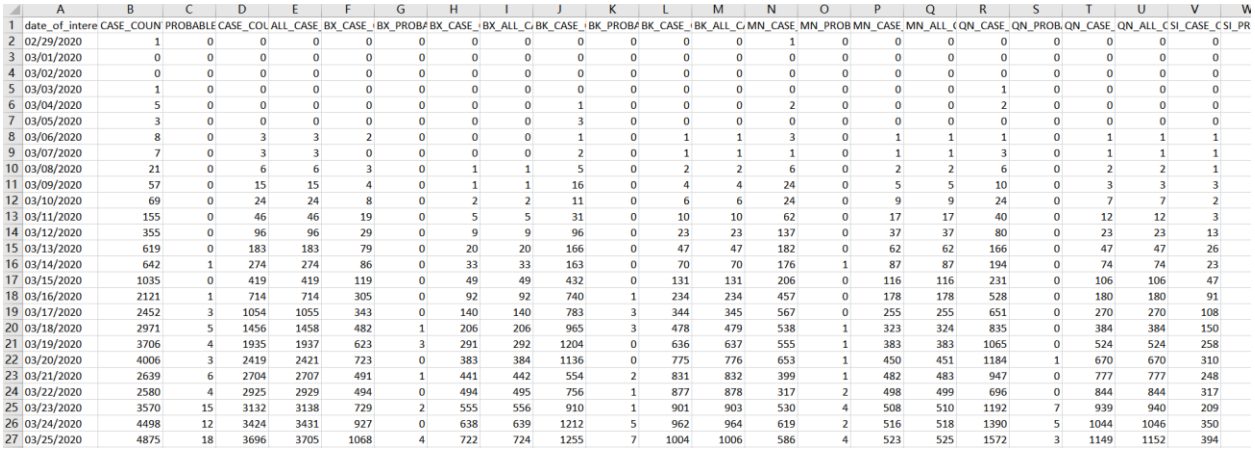
Figure 5. Bounding rectangles for New York, Los Angeles, Toronto, Chicago, Houston, Phoenix, Philadelphia, San Antonio, SanDiego, and Dallas.

	Top Left	Bottom Right	Tweet Count	Percentage (%)
New York	(41.415634, −74.485085)	(40.411124, −71.853181)	20,979	40.3163

Table 1. Coordinates (lat, long) of bounding rectangles for New York city, along with tweet counts and percentages.

3.2. Daily COVID-19 Confirmed Cases in New York City

I obtained the dataset of confirmed COVID-19 cases in New York City for each day in 2020 through the official website, NYC Health (https://www1.nyc.gov/site/doh/covid/covid-19-data-totals.page). The COVID-19 confirmed cases dataset for New York City includes the number of confirmed cases per day and the corresponding 7-day moving average for the entire city, the number of confirmed cases per day for the five boroughs, and the corresponding 7-day moving average for the five boroughs, starting from 2020.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	date_of_intere	CASE_COUN	PROBABLE	CASE_COU	ALL_CASE_	BX_CASE_	BX_PROBA	BX_CASE_	BX_ALL_C	BK_CASE_	BK_PROBA	BK_CASE_	BK_ALL_C	MN_CASE_	MN_PROB	MN_CASE_	MN_ALL_C	QN_CASE_	QN_PROB	QN_CASE_	QN_ALL_C	SI_CASE_C	SI_PR
2	02/29/2020	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
3	03/01/2020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	03/02/2020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	03/03/2020	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
6	03/04/2020	5	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	2	0	0	0	0	
7	03/05/2020	3	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	
8	03/06/2020	8	0	3	3	2	0	0	0	1	0	1	1	3	0	1	1	1	0	1	1	1	
9	03/07/2020	7	0	3	3	0	0	0	0	2	0	1	1	1	0	1	1	3	0	1	1	1	
10	03/08/2020	21	0	6	6	3	0	1	1	5	0	2	2	6	0	2	2	6	0	2	2	1	
11	03/09/2020	57	0	15	15	4	0	1	1	16	0	4	4	24	0	5	5	10	0	3	3	3	
12	03/10/2020	69	0	24	24	8	0	2	2	11	0	6	6	24	0	9	9	24	0	7	7	2	
13	03/11/2020	155	0	46	46	19	0	5	5	31	0	10	10	62	0	17	17	40	0	12	12	3	
14	03/12/2020	355	0	96	96	29	0	9	9	96	0	23	23	137	0	37	37	80	0	23	23	13	
15	03/13/2020	619	0	183	183	79	0	20	20	166	0	47	47	182	0	62	62	166	0	47	47	26	
16	03/14/2020	642	1	274	274	86	0	33	33	163	0	70	70	176	1	87	87	194	0	74	74	23	
17	03/15/2020	1035	0	419	419	119	0	49	49	432	0	131	131	206	0	116	116	231	0	106	106	47	
18	03/16/2020	2121	1	714	714	305	0	92	92	740	1	234	234	457	0	178	178	528	0	180	180	91	
19	03/17/2020	2452	3	1054	1055	343	0	140	140	783	3	344	345	567	0	255	255	651	0	270	270	108	
20	03/18/2020	2971	5	1456	1458	482	1	206	206	965	3	478	479	538	1	323	324	835	0	384	384	150	
21	03/19/2020	3706	4	1935	1937	623	3	291	292	1204	0	636	637	555	1	383	383	1065	0	524	524	258	
22	03/20/2020	4006	3	2419	2421	723	0	383	384	1136	0	775	776	653	1	450	451	1184	1	670	670	310	
23	03/21/2020	2639	6	2704	2707	491	1	441	442	554	2	831	832	399	1	482	483	947	0	777	777	248	
24	03/22/2020	2580	4	2925	2929	494	0	494	495	756	1	877	878	317	2	498	499	696	0	844	844	317	
25	03/23/2020	3570	15	3132	3138	729	2	555	556	910	1	901	903	530	4	508	510	1192	7	939	940	209	
26	03/24/2020	4498	12	3424	3431	927	0	638	639	1212	5	962	964	619	2	516	518	1390	5	1044	1046	350	
27	03/25/2020	4875	18	3696	3705	1068	4	722	724	1255	7	1004	1006	586	4	523	525	1572	3	1149	1152	394	

Figure 5. Daily COVID-19 confirmed cases in New York City data file

3.3 Cleaning Data

After finding, downloading, and understanding the dataset, I began to organize and clean the data on the dataset created by Sara Melotte and Mayank Kejriwal. First, I use a Python script to convert the dataset from a JSON file to a CSV file.

	A	B	C	D	E	F	G	H
1	id	sentiment	date	hashtags	city	state	place_type	
2	1.2407E+18	-0.285416667	Mar 19 2020	[]	New York	SA	admin	
3	1.2407E+18	-0.14875	Mar 19 2020	['CoronaVirus', 'NYC', 'Harlem', 'T	Manhattan	NY	city	
4	1.2407E+18	-0.477777778	Mar 19 2020	[]	Bronx	NY	city	
5	1.2407E+18	0.357575758	Mar 19 2020	['trojan', 'condom', 'nyc', 'corona	Manhattan	NY	city	
6	1.2407E+18	0	Mar 19 2020	[]	Queens	NY	city	
7	1.2408E+18	0.136363636	Mar 19 2020	['corona']	Uniondale	NY	city	
8	1.2408E+18	0.136363636	Mar 19 2020	['corona', 'quarintine', 'damnit', '	Manhattan	NY	city	
9	1.2408E+18	0.136363636	Mar 19 2020	['TBT', 'FIP', 'LowTea', 'HelpMe']	New York	SA	admin	
10	1.2408E+18	0.5	Mar 19 2020	['hiphop', 'rnb', 'pop', 'love', 'rap',	Manhattan	NY	city	
11	1.2408E+18	0.068181818	Mar 19 2020	['chocolatecorona']	Manhattan	NY	city	
12	1.2408E+18	0.6	Mar 19 2020	['nogymnoproblem', 'nogym', 'wo	New Hyde Park	NY	city	
13	1.2408E+18	0.168181818	Mar 19 2020	['corona']	Manhattan	NY	city	
14	1.2408E+18	0	Mar 19 2020	[]	Manhattan	NY	city	
15	1.2408E+18	0.45	Mar 20 2020	['VidaAmericana']	Manhattan	NY	city	
16	1.2408E+18	0.146306818	Mar 20 2020	['corona', 'corona']	Manhattan	NY	city	
17	1.2408E+18	0.05	Mar 20 2020	[]	Jersey City	NJ	city	
18	1.2408E+18	0	Mar 20 2020	['Theater80', 'SaintMarksPlace', '	Manhattan	NY	city	
19	1.2408E+18	0.136363636	Mar 20 2020	['finalcountdown', 'prep', 'paint',	Boonton	NJ	city	
20	1.2408E+18	0.335227273	Mar 20 2020	['barnabynyc', 'barnaby', 'nycdoc	Manhattan	NY	city	
21	1.2408E+18	0.462121212	Mar 20 2020	['covid_19', 'corona', 'coronaviru	New Canaan	CT	city	

Figure 6. The dataset of COVID-19 related tweets posted by people in New York City

I found that some of the data in the dataset were not sorted in date order, they were sorted in a mixed order. So, after sorting all the data in increasing date order, I averaged all the sentiment scores with the same date and obtained the corresponding 7-day moving average. Then I placed the date and corresponding sentiment scores in a separate file. After unifying the date format of the New York City Twitter data with the New York City COVID-19 confirmed data, I placed the daily number of confirmed cases and the corresponding 7-day moving average in the same csv file based on the dates in the Twitter data.

	A	B	C	D	E
1	Date	Sentiment_Avg	Sentiment_MovAvg	Case_Count	Case_Count_7Day_Avg
2	2020/3/19	0.091621989	0	3706	1935
3	2020/3/20	0.156854095	0	4006	2419
4	2020/3/21	0.073128168	0	2639	2704
5	2020/3/22	0.119493414	0	2580	2925
6	2020/3/23	0.145771283	0	3570	3132
7	2020/3/24	0.102612237	0	4498	3424
8	2020/3/25	0.129227978	0.119583111	4875	3696
9	2020/3/26	0.115168622	0.119818361	5044	3887
10	2020/3/27	0.110882867	0.11316359	5119	4046
11	2020/3/28	0.189448836	0.129058688	3479	4166
12	2020/3/29	0.136965774	0.131884929	3560	4306
13	2020/3/30	0.199569239	0.138490086	6129	4672
14	2020/3/31	0.179156333	0.149985335	5458	4809
15	2020/4/1	0.135558367	0.151776963	5449	4891
16	2020/4/2	0.150837247	0.159037809	5747	4992
17	2020/4/3	0.057033296	0.151692224	5669	5070
18	2020/4/4	0.159974279	0.146771176	3864	5125
19	2020/4/5	0.185936334	0.151807324	3780	5157
20	2020/4/6	0.166762065	0.146887221	6353	5189
21	2020/4/7	0.032515674	0.13104967	6043	5272
22	2020/4/8	0.238335946	0.14134954	5576	5290
23	2020/4/9	0.09538033	0.133352311	5071	5194
24	2020/4/10	0.149703251	0.148683737	4510	5028
25	2020/4/11	0.122229324	0.142183734	3733	5009
26	2020/4/12	0.154740641	0.138212029	2888	4882
27	2020/4/13	0.140822999	0.134330872	3311	4447

Figure 6. Cleaned data file

3.4 Calculating Correlation

In this study, I used the quantitative method of correlation analysis to determine the relationship between the sentiment scores of COVID-19-related tweets posted by people in New York City and the daily case confirmations of people in New York City beginning in March 2020. The use of correlation to determine the relationship and the strength of the relationship is a classic quantitative approach [25]. Correlation analysis measures the degree of association between two variables [26]. Correlation can be positive or negative.

- Positive correlation: two variables change in the same direction.
- Neutral correlation: these variables are uncorrelated, and their changes are unrelated.
- Negative correlation: the variables change in the opposite direction.

There are two commonly used correlation methods, Pearson's Correlation and Spearman's Correlation.

3.4.1 Pearson's Correlation

Pearson's Correlation is used to summarize the strength of the linear relationship between two data samples. If the correlation coefficient is positive, it indicates that the value of one variable is associated with the value of the second variable [26]. The calculation of Pearson's Correlation between two variables is defined as the product of the covariance of the two variables divided by the standard deviation of each data sample. It normalizes the covariance between the two variables and then gives an interpretable score.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

Figure 7. Formula for calculating the Pearson's correlation coefficient between two variables

Pearson's Correlation is a value between -1 and 1 [26] and represents the limit of correlation from a perfectly negative correlation to a perfectly positive correlation. If the value is 0, then it means that there is no correlation. According to Quinnipiac University's strength of correlation scale, when the Pearson‘s correlation coefficient is greater than or equal to 0.7, the correlation is considered very strong positive; when the Pearson correlation coefficient is less than or equal to -0.7 , the correlation is considered to be a very strong negative correlation; when the Pearson correlation coefficient is less than or equal to 0.2, the correlation is considered to be a weak positive correlation; and when the Pearson correlation coefficient is greater than or equal to -0.2, the correlation is considered to be a weak negative correlation.

The Pearson's correlation coefficient is one of the commonly used methods, however, it is important to note whether the data set satisfies the necessary conditions of Pearson's correlation coefficient.

- There is a linear relationship between the two variables.
- The variables are continuous variables.
- The variables both conform to a normal distribution and their binary distribution also conforms to a normal distribution.
- The relationship between the two variables is independent.
- The variance of the two variables is not 0.

As we can observe from the Figure 8 and Figure 9, these four variables are not actually normally distributed. Therefore, the Pearson’s correlation is not a suitable method for the analysis of these four variables.

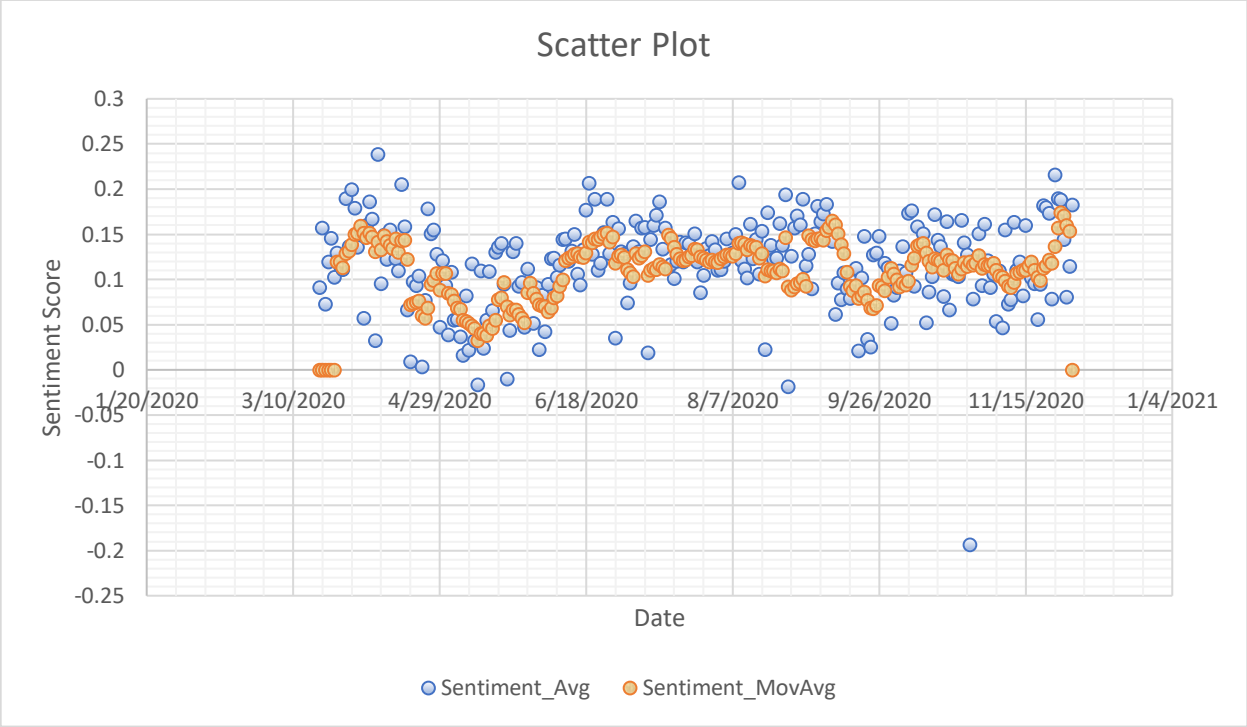


Figure 8. Scatter plot of sentiment average and 7-day sentiment moving average

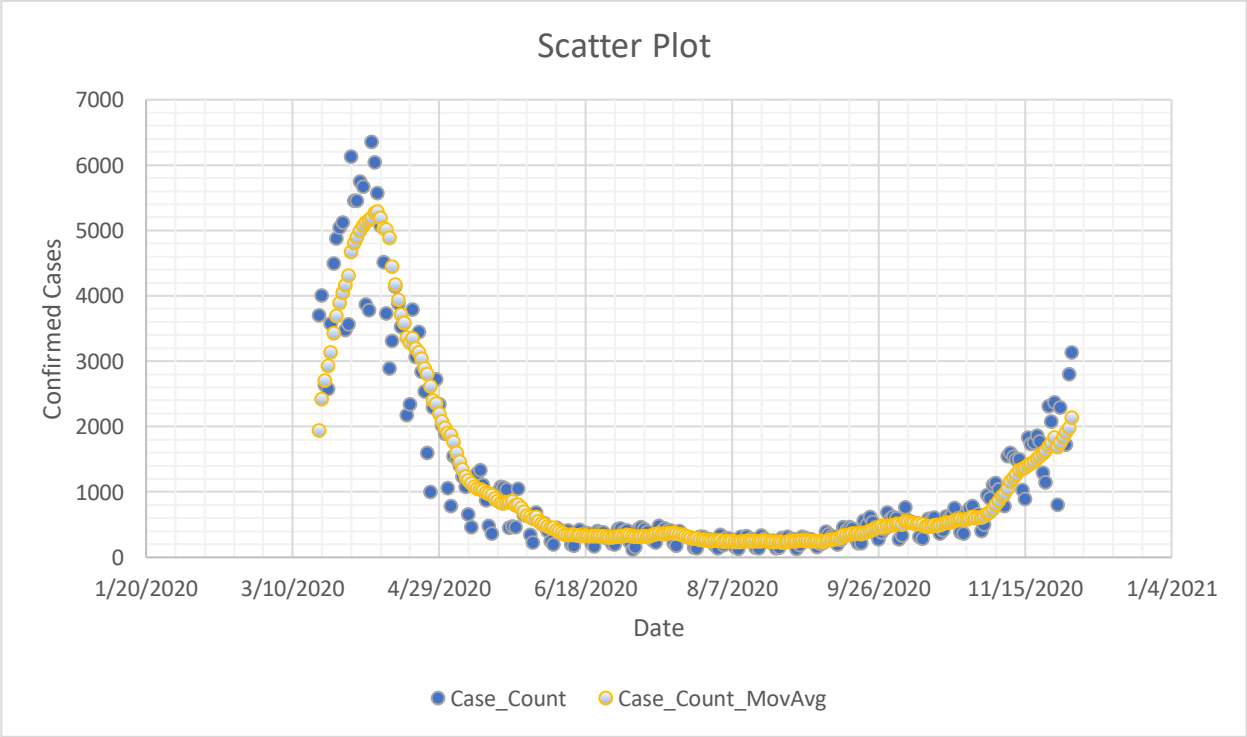


Figure 9. Scatter plot of daily confirmed cases and 7-day confirmed cases moving average

3.4.2 Spearman’s Correlation

Another commonly used correlation method is Spearman's Correlation, which is known as one of the three major statistical correlation coefficients along with Pearson's and Kendall's correlation coefficients. Spearman's Correlation can be used wherever the Pearson Correlation can be used, unless performance effects are taken into account. The difference between these two correlation coefficients is that the Pearson’s correlation measures the strength of linear correlation between two normally distributed variables, while the Spearman’s correlation measures the relationship between. Therefore, Spearman's correlation coefficient has a wider range of applications, not only because it does not make any assumptions about the data distribution, but also because it tolerates outliers and does not require the data to be equally spaced.

Spearman's Correlation is used to measure the strength of the monotonic relationship between two continuous variables. It is also a value between -1 and 1 and represents the limit of correlation from a perfectly negative correlation to a perfectly positive correlation. In the absence of repeated data, if one variable is a strictly monotonic function of the other, the Spearman's correlation coefficient is either 1 or -1. If the correlation is positive, this means that the two variables are moving in the same direction. If the correlation is negative, it means that when the value of one variable increases, the value of the other variable decreases. When the correlation is neutral or zero, it means that these variables are uncorrelated.

The Spearman's correlation coefficient is presented as ρ_s. If each variable does not have the same value, the Spearman’s correlation coefficient can be calculated by the following equation.

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Figure 10. Formula for calculating the Spearman's correlation coefficient between two variables

In this formula, n is the number of data points, and d_i is the difference of the rank order (rx_i,ry_i) of the data points (x_i,y_i) : $di=rx_i-ry_i$.

If a variable has repeated data, the calculation of the Spearman's correlation coefficient between the variables is the calculation of the Pearson's correlation coefficient between the data ranks of the variables.

$$\rho_s = \rho_{r_x, r_y} = \frac{\operatorname{cov}(r_x, r_y)}{\sigma_{r_x} \sigma_{r_y}}$$

Figure 11. Formula for calculating the Spearman's correlation coefficient between two variables when a variable has repeated data

r_x is the rank of the transformed variable x. From this definition, it can be observed that the Spearman's correlation coefficient is the Pearson's correlation coefficient after the rank transformation of the data. When the Pearson's correlation is large, the Spearman's correlation is also large; while when the Pearson's correlation is small, the Spearman 's correlation may still be larger. This means that when there are outliers in the data set, their Pearson's correlation is more affected, but Spearman's correlation is more tolerant of outliers.

4.Result

4.1 Sentiment Analysis

The dataset for this study collected a total of 20,980 COVID-19-related tweets posted by people in New York City and the corresponding daily confirmed COVID-19 cases, starting on March 20, 2020.

From the table 2, we can observe that there is a difference between the maximum and minimum values of Sentiment Score Average and Sentiment Score 7-Day Moving Average, but their mean values are very close to each other. The average of COVID-19 Confirmed Cases and the average of COVID-19 Confirmed Cases 7-Day Moving Average are very different, and their mean values are also very different. However, the dates on which their maximum and minimum values occur are relatively close to each other.

	Max Value	Min Value	Average	Date
Sentiment Score Average	0.238335946	-0.193333333	0.159350659	Max: 2020/04/08 Min: 2020/10/27
Sentiment Score 7-Day Moving Average	0.188045852	0.032355817	0.157861295	Max: 2020/11/27 Min: 2020/05/12

COVID-19 Daily Confirmed Cases	6353	120	2132.285714	Max: 2020/04/06 Min: 2020/07/04
COVID-19 Daily Confirmed Cases 7-Day Moving Average	5290	229	1874.428571	Max: 2020/04/08 Min: 2020/08/25

Table2. Statistical results of the dataset

The Figure12 and Figure13 show the trend of sentiment scores and the number of confirmed cases from March 19, 2020, to December 01, 2020, respectively. From the trend graph of the sentiment scores, we can observe that from March to May 2020, the sentiment scores show some decreasing trend. From May to June 2020, the sentiment scores show some upward trend. Moreover, most of the data for sentiment scores are positive. From the graph of confirmed cases, we can observe that from March to April 2020, the number of confirmed cases per day in New York City shows an increasing trend, from April to June shows a decreasing trend, from June to October shows a flat trend, and then from October to December shows an increasing trend again.

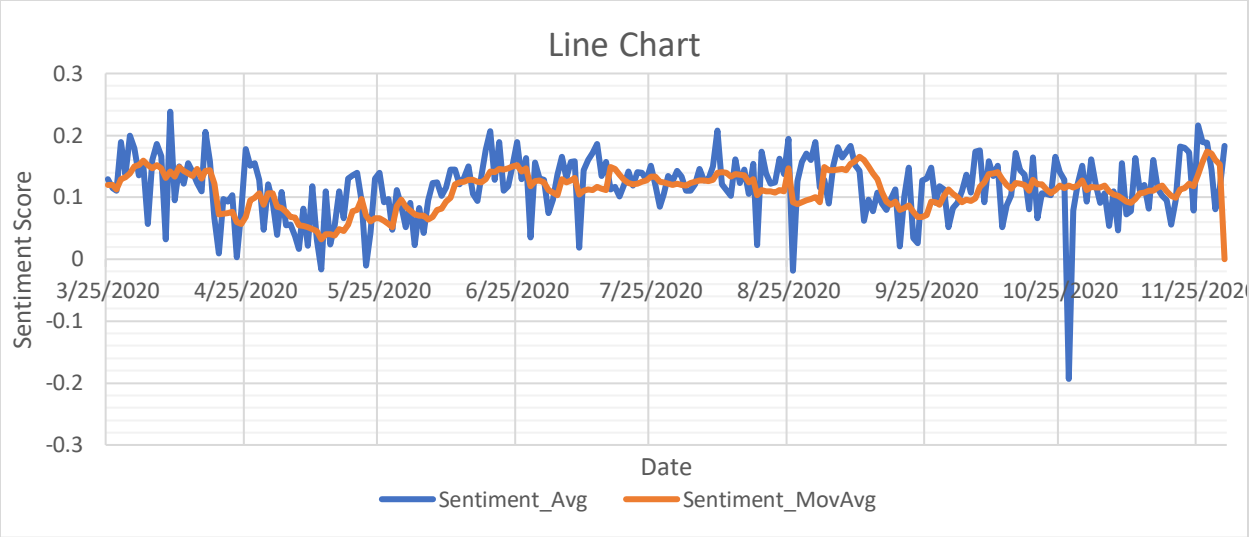


Figure 12. Line chart of sentiment Scores

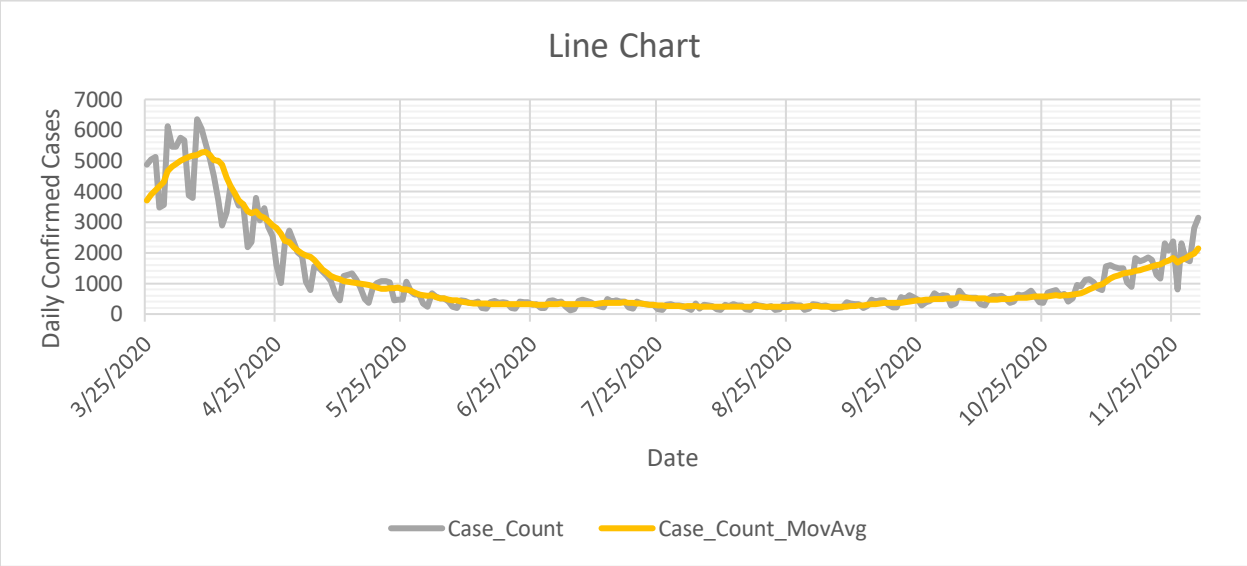


Figure 13. Line chart of daily confirmed cases

4.2 Correlation Calculation

Table 3 shows the results of Spearman's Correlation coefficients and the corresponding P values for the four variables through a script in Python.

	Spearman's Correlation	P-value
7-Day Moving Average of Sentiment Score and 7-day Moving Average of Confirmed Cases	-0.20874925880535203	0.0008556856555061927
7-Day Moving Average of Sentiment Score and Daily Confirmed Cases	-0.13513673622650677	0.03200004946476612
Average of Sentiment Score and 7-day Moving Average of Confirmed Cases	-0.1556094966251999	0.012329011716778632
Average of Sentiment Score and Daily Confirmed Cases	-0.15267580163667338	0.014094843362271238

Table 3. Result of Spearman's Correlation coefficients and the corresponding P values for the four variables

5. Discussion

5.1 Relationship between Sentiment Scores and Confirmed Cases

Combining the trend of sentiment scores of Tweets about COVID-19 posted by people in New York City collected from March to December 2020 and trend of daily confirmed cases in New York City, we can observe that,

- from May to June, people's sentiment scores show an increasing trend, while daily confirmed cases show a decreasing trend.
- from July to August, people's sentiment scores do not have many fluctuations, and daily confirmed cases show a relatively flat trend.
- from August to September, people's sentiment scores show a decreasing trend, while daily confirmed cases show a relatively flat upward trend.
- from October to November, people's sentiment scores show a relatively downward trend, while daily confirmed cases show a relatively upward trend.

During these overlapping time periods, sentiment scores and confirmed cases show some opposite trends.

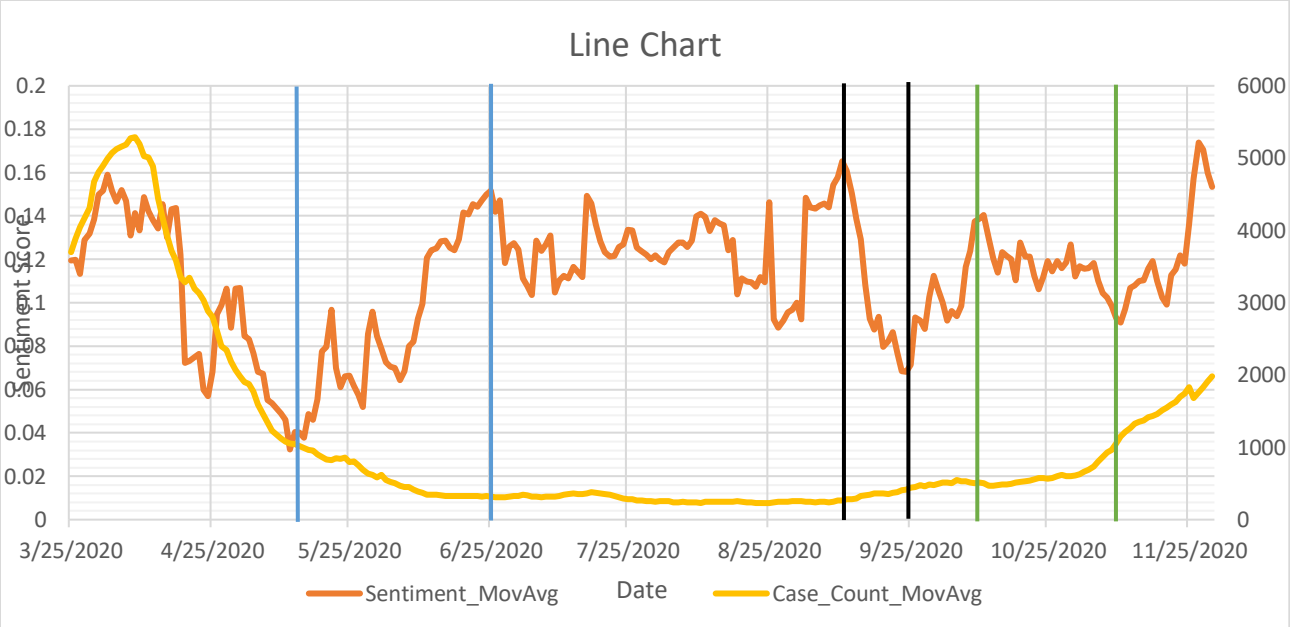


Figure 13. Line chart of 7-day sentiment moving average and 7-day confirmed cases

Spearman's Correlation is used to summarize the monotonic relationship between variables. It has a value between -1 and 1 and represents the limit of correlation from a perfectly negative correlation to a perfectly positive correlation. If the value is 0, then it means that there is no correlation. From the calculation of Spearman's Correlation in Table, we can observe that the relationship between the 7-day average, 7-day moving average of the sentiment scores of Tweets about COVID-19 posted by people in New York City, daily confirmed cases and 7-day moving average of confirmed cases in New York City is negative. This relationship is statistically significant between sentiment and daily cases with $p < 0.05$. And the four p-value showed in Table is less than 0.05. Thus, such a negative relationship could suggest that people's emotions are affected by the COVID-19 epidemic in New York. The negative correlation between sentiment scores and daily confirmed cases, combined with the trend of them showed in Figure, we could suggest that the trend of people's sentiment and daily confirmed cases changed in opposite ways during certain time periods. When the number of confirmed cases is decreasing, there is an upward

trend in people's mood scores (people become relatively positive). However, when the number of confirmed cases increased, the sentiment scores tended to decrease (people become relatively negative).

5.2 Limitation

However, this study has some limitations in the data set. Although I obtained a negative relationship between the sentiment scores of tweets posted by people in New York City and confirmed cases per day in New York City based on Spearman's Correlation, and all the p-value are less than 0.05, indicating that such a negative relationship is statistically significant. These correlation coefficients are not high, which implies that there is not a very strong negative relationship between these variables. Moreover, the average of all sentiment scores in the dataset is positive, which indicates that the average sentiment of people during the 255-day period of collected NYC tweets is positive. However, it is obvious that in North America, people did not feel positive about the crisis in general.

The data I downloaded on COVID-19 tweets is a subset of the GeoCOV19Tweets dataset. Two researchers used tools to determine more specific location information in the GeoCOV19Tweets dataset and then retrieved more detailed Twitter data about New York City. As explained in this article, their dataset retains the sentiment scores from the GeoCOV19Tweets dataset.

The GeoCOV19Tweets dataset uses TextBlob to get all the sentiment scores that are available. For most natural language processing projects with "normal" text, such as books, news articles, movie reviews, etc., we can usually use TextBlob for sentiment analysis. TextBlob is a library that provides a simple API to process text data with tasks such as part-of-speech data, noun phrase extraction, tokenization, classification, etc. TextBlob is unique for sentiment analysis because, in

addition to polarity scores, it can also generate subjectivity scores. However, Sidney Kung found in her research that the performance of TextBlob on Twitter data is not representative [27]. The sentiment scores derived from analyzing Twitter via TextBlob do not appear to be representative for some of the tweets in her Twitter dataset. For example, TextBlob is not sensitive to analysis of hate speech and offensive language. As a result, TextBlob might encounter difficulties in analyzing Twitter data. Therefore, it is said that TextBlob may not be the most appropriate tool for sentiment analysis of social media. The low strength of correlation between the two variables may have the reason of imprecise sentiment score.

6. Conclusion

Based on the negative Spearman's Correlation between sentiment scores of COVID-19 related Tweets posted by people in New York City and daily confirmed cases, we could suggest that the relationship between people's emotion during the COVID-19 and the spread of the pandemic is opposite. People's moods would get worse as COVID-19 gets worse, and they would get positive as COVID-19 gets better.

However, the negative coefficient is greater than -0.2, which means that the strength of the negative relationship is weak. The tools used to perform sentiment analysis may have contributed to this weak relationship result. There are some differences between sentiment analysis of social media and sentiment analysis of other texts. For example, when analyzing sarcastic sentiment in social media, the analysis tool needs to have a good sensitivity to sarcasm and be able to correctly distinguish objects in the text. This is key to obtaining more accurate sentiment scores. For example, the same text will express different emotions when describing different objects. When we describe a concert as crazy, we have a positive attitude toward the concert. But if we say that the political leader is crazy, then the expression contains a negative attitude. Therefore, it is very important to discern negative attitudes when performing sentiment analysis in social media. For example, in this study, TextBlob, the tool used to perform sentiment analysis, is not sensitive to negative terms in social media. This may result in the analyzed sentiment being more positive than the actual sentiment. Although there are many sentiment analysis tools available, we need to be aware of the ability of these tools to identify negative words when performing sentiment analysis on social media.

7. Reference

[1] H. Wang et al., “Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China,” Cell Discovery, vol. 6, no. 1, pp. 1–8, Dec. 2020.

[2] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in Twitter,” in Proc. ICCC, 2014, pp. 155–162.

[3] M. J. Widener and W. Li, “Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US,” Appl. Geography, vol. 54, pp. 189–197, Oct. 2014.

[4] Ahmed W. Using Twitter as a data source: an overview of social media research tools (updated for 2017)[J]. Impact of Social Sciences Blog, 2017.

[5]. Carter M. How Twitter may have helped Nigeria contain Ebola. BMJ 2014 Nov 19;349:g6946. [doi: 10.1136/bmj.g6946] [Medline: 25410185]

[6]. Dalrymple KE, Young R, Tully M. “Facts, Not Fear”: negotiating uncertainty on social media during the 2014 Ebola crisis.Sci Commun 2016 Jun 22;38(4):442-467. [doi: 10.1177/1075547016655546]

[7]. Guidry JP, Jin Y, Orr CA, Messner M, Meganck S. Ebola on Instagram and Twitter: how health organizations address the health crisis in their social media engagement. Public Relations Rev 2017 Sep;43(3):477-486. [doi:10.1016/j.pubrev.2017.04.009]

[8]. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control 2015 Jun;43(6):563-571.[doi: 10.1016/j.ajic.2015.02.023] [Medline: 26042846]

[9]. Rufai S, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. J Public Health (Oxf) 2020 Aug 18;42(3):510-516 [FREE Full text] [doi: 10.1093/pubmed/fdaa049] [Medline: 32309854]

[10]. Bakshi R K, Kaur N, Kaur R, et al. Opinion mining and sentiment analysis[C]//2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, 2016: 452-455.

[11]. Liu B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1): 1-167.

[12]. ComScore/the Kelsey group, “Online consumer-generated reviews have significant impact on offline purchase behavior,” Press Release, http://www. comscore.com/press/release.asp?press=1928, November 2007.

[13]. Tumasjan A, Sprenger T, Sandner P, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2010, 4(1).

[14]. W. He, H. Wu, G. Yan, V. Akula, and J. Shen, “A novel social media competitive analytics framework with sentiment benchmarks,” Inf. Manage., vol. 52, no. 7, pp. 801–812, Nov. 2015.

[15]. Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

[16]. Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]//Lrec. 2010, 10(2010): 2200-2204.

[17]. Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining[C]//Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). 2006.

[18]. Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. Proceedings of @NLP can u tag #usergeneratedcontent?!, 15-22.

[19]. Gruzd, A.; Mai, P. COVID-19 Twitter Dataset; Scholars Portal Dataverse: Vancouver, Canada, 2020, doi:10.5683/SP2/PXF2CU. [CrossRef]

[20]. Chen, E.; Lerman, K.; Ferrara, E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health Surveill 2020, 6, e19273. [CrossRef] [PubMed]

[21]. Qazi, U.; Imran, M.; Ofli, F. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. Sigspatial Spec. 2020, 12, 6–15. [CrossRef]

[22]. Baran, E.; Dimitrov, D. TweetsCOV19-A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. 2020. Available online: https://dl.acm.org/doi/abs/10.1145/3340531.3412765 (accessed on 15 June 2021).

[23]. Melotte S, Kejriwal M. A Geo-Tagged COVID-19 Twitter Dataset for 10 North American Metropolitan Areas over a 255-Day Period[J]. Data, 2021, 6(6): 64.

[24]. Lamsal, R. Design and analysis of a large-scale COVID-19 tweets dataset. Appl. Intell. **2020**, 51, 2790–2804. [CrossRef]

[25]. Byrne, G. (2007). A statistical primer: Understanding descriptive and inferential statistics. *Evidence Based Library and Information Practice, 2*(1), 32-47.

[26]. Moutinho, L. (2011). Correlation analysis. In L. Moutinho, & G. Hutcheson (Eds.), The SAGE dictionary of quantitative management research. (pp. 57-61). London: SAGE Publications Ltd. doi: http://dx.doi.org/10.4135/9781446251119.n17

[27]. Sidney Kung(2021). Social Media Sentiment Analysis with VADER, Measuring the sentiment for nuanced text data. https://towardsdatascience.com/social-media-sentiment-analysis-with-vader-c29d0c96fa90