#COVID-19 Searching for a Relationship between Twitter Sentiment and Infectious Disease

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

James Odysseus Van Der Loo Stassinos
Bachelor of Arts
Virginia Polytechninic Institute and State University, 2019

Director: Dr. Taylor Anderson, Professor
Department of Geography and Geoinformation Science

Spring Semester 2023
George Mason University
Fairfax, VA

Dedication

I dedicate this thesis to Mom and Dad.

Acknowledgments

This study was funded by the National Science Foundation (NSF) Awards #2109647, #2302968, and #2302970.

Data availability statement

Data processing scripts are available through the following github repository `https://github.com/jstassinos/Processing-the-TBCOV-Dataset`. The TBCOV geolocated Tweet dataset is located at this website for download `https://crisisnlp.qcri.org/tbcov`. All hydrated Tweets data is available through the Twitter API with a Twitter Developer License. COVID-19 aggregated case data is available on the USA facts `https://static.usafacts.org/public/data/covid-19/covid_confirmed_usafacts.csv`

Table of Contents

List of Tables

List of Figures

Abstract

#COVID-19 SEARCHING FOR A RELATIONSHIP BETWEEN TWITTER SENTIMENT AND INFECTIOUS DISEASE

James Odysseus Van Der Loo Stassinos

George Mason University, 2023

Thesis Director: Dr. Taylor Anderson

Digital health data such as social media data has shown potential for identifying outbreaks faster than official records of disease incidence. The objective of this thesis [1] is to examine the relationship between COVID-19-related Tweet sentiment and COVID-19 cases over space and time and assess the extent to which Twitter-derived sentiment can be used for local COVID-19 surveillance in the United States. To our knowledge, there is no existing study that examines the relationship between Tweet sentiment and infectious disease cases at a spatially local level. The sentiment is computed using 56,755,894 Tweets from the TB-COV dataset for US counties over time. Tweet sentiment is examined with COVID-19 cases for each county globally, over time, and space using Pearson's R correlation. A negative association was observed between COVID-19 cases and the sentiment polarity of COVID-19 tweets, but only in some regions of the US and only for some duration of the period of study. Further research is needed to understand the cause of the spatial and temporal non-stationary correlations between Twitter sentiment and COVID-19 cases. This would allow for the identification of when and where Twitter sentiment could be used as a signal for early disease outbreak warning.

[1] The original thesis research has been submitted to Nature Scientific Reports under the title "Towards using Twitter Sentiment for Infectious Disease Detection" and is currently under review

Chapter 1: Introduction

Accurate and timely data capturing the trajectory of COVID-19 pandemic has been crucial for informing policy interventions [1, 2]. To measure the cases and deaths resulting from SARS-CoV-2 across the US, the Center for Disease Control (CDC) gathers data from jurisdictional and state partners that independently report the new daily number of confirmed COVID-19 cases and deaths for each county of residence based on testing data captured by pharmacies, hospitals, and other testing facilities [3]. However, due to a decentralized and fragmented public health infrastructure, this data is subject to inherent geographic biases, a lack of standards, and reporting lags [4]. One such example is testing bias, where some locations may have better testing infrastructure, well-funded access to testing, and less stigma around getting tested [4]. In another example, a lack of standardization resulted in an inconsistent set of reporting metrics metrics where some states defined a positive case as a total count of positive tests and others as the total number of unique individuals that tested positive [5]. Furthermore, delays and backlog in the reporting pipeline result in on average a 3-21 day lag between the time a patient is tested positive and reported as having tested positive [6].

Publicly available data from news outlets, chat rooms, web searches, or social media are often used as a supplement to official data sources to identify outbreaks of existing or emerging diseases faster than what is reported by the CDC and can even detect outbreaks not detected by official sources [7–10]. Collectively, these sources, referred to as digital health data, provide a lens into public health that is fundamentally different from that yielded by official sources [11]. Among such sources, Twitter is used to detect the prevalence of diseases such as influenza and dengue fever. In some cases, the volume of tweets is used as an indicator for disease. More specifically, the number of tweets that contain keywords

relating to a disease or have been classified as a "self report" have been found to correlate with CDC reported cases and other official sources and thus can be used as a tool for early detection and monitoring [12–17].

In other cases, additional analyses of the tweets, using natural language processing approaches such as sentiment analysis, has been used to differentiate between tweets that are positively i.e. "my flu shot worked, no flu for me!" and negatively associated with disease i.e. "my whole family has the flu" and predict to what extent influenza is present in the population over time [18–21]. Where tweet sentiment and official influenza cases tend to be inversely associated, the relationship is less clear for COVID-19. For example, Valdez et al [22] were surprised to observe a positive correlation between US wide COVID-19 related tweet sentiment and cases and deaths, meaning that as cases and deaths increase in the US, sentiment towards COVID-19 trends positive. This contradicts what would intuitively be expected. In another example, Feng and Kirkley [23] find a weak negative or absent correlation between state-level Twitter sentiment and COVID-19 and cases and deaths. They use the GEOCOV19 dataset [24] which contains geolocated tweets spanning only Feb 1, 2020 to May 1, 2020 in 49 cities across the United States and is the precursor dataset to the larger TBCOV dataset [25]. Given the limited and conflicting number of studies, there is a need to thoroughly examine the relationship between tweet sentiment and COVID-19 cases across space and time and the potential to use such information as a surveillance tool.

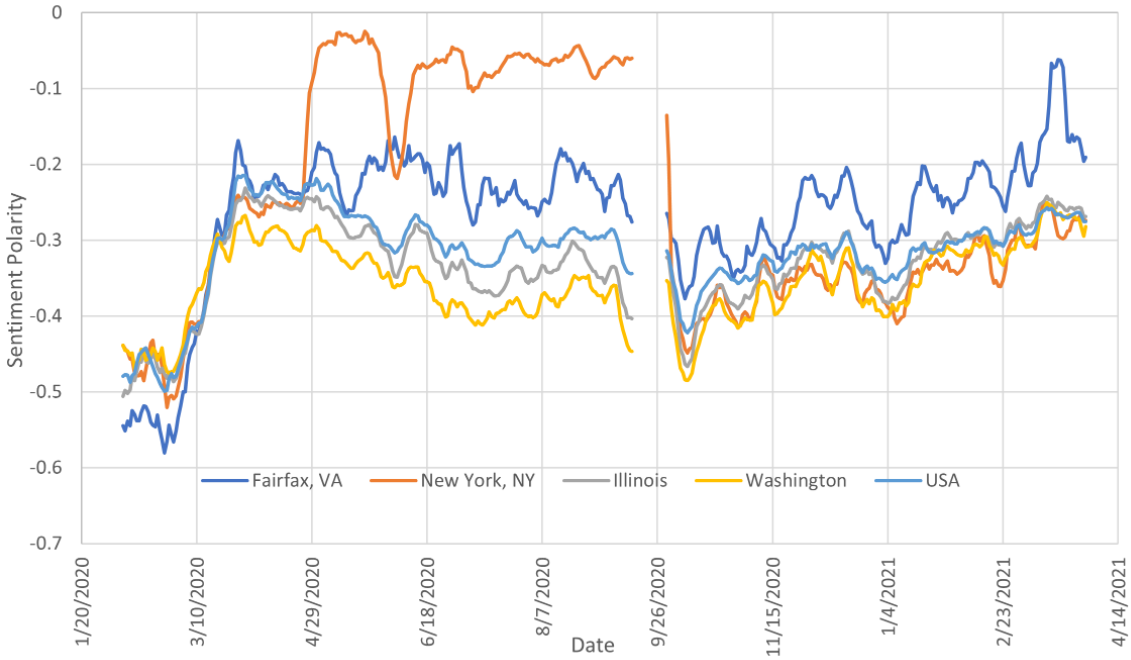
Therefore, this thesis seeks to answer the following research questions:

1. What is the relationship between COVID-19 related tweet sentiment and official case data over space and time?

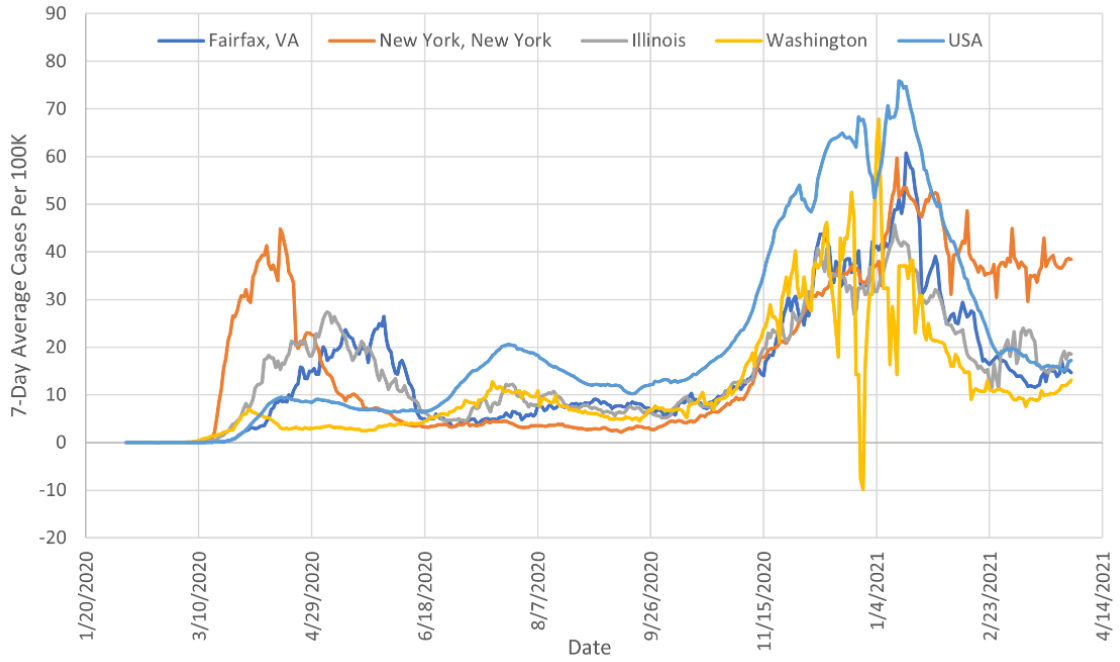
2. What is the extent to which Twitter derived sentiment can be used for local COVID-19 surveillance in the United States?

The thesis hypothesizes that there should be a strong negative association between local sentiment and cases, where as cases in a county go up, county sentiment goes down (and vice versa). The thesis also hypothesizes that this relationship will be the same across space and

time, also known as spatial and temporal stationarity [26]. Towards answering these research questions and testing these hypotheses, this thesis leverages the full TBCOV dataset [25], containing a total of two billion geolocated COVID-19 related tweets between Jan 2020 to Dec. 2021. The county level correlation between tweet sentiment and official data from February 1, 2020 to March 31, 2021 is examined in order to determine whether, when, and where tweet sentiment from a county can be used as a predictor of the number of cases in the same county.



(a) Tweet Sentiment.



(b) New COVID-19 Cases.

Figure 1.1: Seven-day Rolling average of COVID-19 sentiment and new COVID-19 cases.

Intuition would suggest that there should be an inverse association between COVID-19 tweet sentiment and COVID-19 cases at the local level. Thus, places experiencing a high number of COVID-19 cases are expected to have a low COVID-19 sentiment during that time. This inverse association can observed in Figure 1.1 which shows both the COVID-19 related sentiment on Twitter and the number of COVID-19 cases for the entire U.S. and for some locations in the U.S. For example, in the case of New York City, near the end of April, when COVID-19 cases dramatically decrease, one can observe a large increase in mean sentiment. Yet, in general it is difficult to discern any clear relationship, thus, warranting further investigation. In the case that there is an inverse association between COVID-19 tweet sentiment and COVID-19 cases at a local level, it is anticipated that COVID-19 related tweet sentiment could be used to supplement official disease surveillance data streams and provide important insights for local outbreak detection of diseases [27]. As far as is currently known, there is no existing study that examines the relationship between tweet sentiment and infectious diseases cases at a spatially local level.

This thesis is organized as follows: Chapter 2 introduces social media as a proxy for disease surveillance. Chapter 3 describes the tweet dataset used and any pre-processing methods used. Chapter 4 describes the experiments and how tweet sentiment is compared to COVID-19 cases. Chapter 5 presents the results and emphasizes noteworthy information. Chapter 6 provides context for the results and how tweet sentiment correlates with cases. Chapter 7 describes how this research contributes to disease surveillance using Twitter sentiment as a data source.

Chapter 2: Related Work

Infectious disease outbreak detection has come a long way since John Snow's revolutionary discovery that contaminated water was causing Cholora outbreaks in London[28]. In modern times predicting disease incidence using publicly available data from news outlets, chat rooms, web searches, or social media data has become an increasingly popular research topic, given the vast amounts of health-related information that people share on these platforms. This type of data is referred to as digital health data and has the potential to complement traditional disease surveillance methods and provide timely and cost-effective insights for mitigating the spread of illnesses. However, the use of digital health data for disease prediction is a relatively new area of investigation.

2.1 Digital Health Data for Disease Detection

In general, digital health data has been shown to be a viable method for detecting diseases. Ginsberg et al. [29] use the weekly volume of influenza related web searches normalized by the total number of searches to estimate the level of weekly influenza in each region of the United States. They compare their search volume to the CDC reported Influenza Related Illness percentages for an average correlation of 0.90 between 2003 and 2007. Blogging is another form of digital health data that has shown positive results in detecting flu like illnesses. For example, Corley et al. [30] was able to detect the beginning of the 2008 flu season using blogs. They count the number of blogs that contain keywords mentioning flu like illnesses comparing this to the CDC's Outpatient Influenza-like-illness Surveillance Network (ILINet) and produce a Pearson's correlation of r = 0.767 with 95% confidence. In another interesting example, Beauchamp [31] explains that detecting the loss of smell through online text can be difficult due to infrequent discussions. Using a unique data source in Amazon

candle reviews, and a novel Bayesian Vector Autoregression determined that COVID-19 cases appear to significantly effect the "no smell" review rates [31]. Alternatively, Human mobility data has been used to understand human behavior in various applications [32]. In the context of understanding infectious diseases, mobility data has also been used for infectious disease spread visualization [33], understanding of spread processes [34, 35] for contact tracing [36], and to inform infectious disease spread simulations [37, 38].

In this section, the existing literature on this topic will be reviewed specifically focusing on the sentiment analysis of Twitter data as a digital health data source, sometimes also referred to as volunteered geographic information (VGI) [39].

2.2 Twitter for Disease Surveillance

Microblogging platforms, specifically Twitter is a valuable source for VGI that can be used for disease detection. Recent research into influenza prediction from micro-blogging platforms has yielded positive results in moving towards a strengthened public health prediction model. Sadilek et al. [40] finds that using Twitter data and their probabilistic model they can predict with high precision and good recall if that person will fall ill. They start with a Support vector machine (SVM) classifier to identify sick users within the corpus. A prediction can be made using a Conditional Random Field (CRF) model that incorporates that user's social network and co-location. The drawback of this study in the current environment is Twitter has restricted the amount of access researchers have to exact coordinates, limiting the repeatability of the co-location portion of the study.

Another method for outbreak detection is proposed in the research of Paul and Dresde [41] showing a correlation coefficient of 0.958 when comparing CDC Flu view data to their Ailment Topic Aspect Model plus's (ATAM+) "flu" ailment Probability that uses Twitter data. The ATAM+ model is based on topic models such as latent Dirichlet allocation (LDA) and is fine tuned for topics relating to how users express their illnesses and ailments. The tweet dataset starts with 2 billion tweets and after classifying all the health related tweets

they have a corpus of 1.63 million English tweets[41]. This research shows positive results, however it needs to be more spatially precise if its to be use full for containing local outbreaks with precision mitigation.

The previous two studies focus on diseases that well known by the public and have been formalized by national public health institutes. Lim et al. [42] proposes a method for detecting latent diseases using tweet data as it can take time for a top down style public health institution to recognize a novel disease which is why a bottom up approach is required. They focus on users over time who tweet about potential symptoms with negative sentiment, this way they can infer that the user is suffering from the symptom. Then symptom weighting vectors are used to match the symptoms from a user's tweets about to their electronic medical records (EMR). As EMR's are private by law they recruited 104 volunteers who were diagnosed with influenza from the Penn State's Health Services to participate in the study. The resulting F1 score of 0.724 shows that this model is use full for detecting latent infectious disease without the use of manually classified data.

Looking outside the United States, in China, Weibo is a popular social media platform similar to Twitter. Shen et al. [43] analyzed COVID-19 related posts and were able to classify "sick posts" based on user's reported symptoms. These "sick posts" are able to predict daily case counts up to 14 days ahead of the public health infrastructure's official statistics. Providing evidence that social media posts relating to diseases can be predictive of epidemic measures.

2.3 Sentiment For Disease Surveillance

Sentiment analysis is a sub field of natural language processing that is concerned with the identification and categorization of opinions in text. Valdez et al. [22] applied this to tweets using Vader Sentiment to calculate polarity. They were surprised to learn that the polarity for the COVID-19 corpus of tweets increased becoming more positive starting when the WHO declared a pandemic on March 10, 2020 to the end of March. In deeper COVID-19 tweet sentiment research Feng and Kirkley [23] compare polarity to epidemic measures

(cases and deaths) and human mobility patterns. They report high associations between online emotional responses and offline mobility but very little correlation with geolocalized mean sentiment values. In the past, research on tweet sentiment for disease detection used sentiment classification in the pre-processing stage to identify tweet sentiment. The classified tweets are then used for analysis, and not raw sentiment values.

There are many papers using tweet sentiment as a proxy for public opinion and topic opinion concerning diseases, but there is limited evidence that sentiment can be used as a proxy for disease incidence detection. The following paragraph describes some examples. Signori et al. [44] used google maps to display tweets as points on a continuously updating web map in order to show the evolving conversation surrounding H1N1. They used the Geolocation of the tweets to estimate Influenza like illness (ILI) rates, however sentiment polarity was not utilized as a basis for comparison with ILI rates in the prediction process. Behexra and Eluri [45] created a Degree of concern (DOC) metric using sentiment analysis for measuring how a population felt about different diseases over time and space. They used a two step sentiment calculation process where a user's personal negative tweets were identified. This method is designed to extract a user's personal level of concern and then aggregate it for the population at large. They do not take the extra step to compare their degree of concern metric against disease incidence levels. Ji et al. [46] used a similar method for sentiment analysis idenfifying personal and non personal negative tweets from Twitter users to create a Measure of Concern (MOC) metric. This MOC metric was unable to be used for prediction but the peaks matched up with peaks in news volume suggesting negativity increases as news increases.

2.4 Summary

There are many papers that are able to find a link between digital health data and disease incidence[29–31]. In digital health data obtained from Twitter, studies have examined the correlation between the volume of tweets often it and CDC ILI data [41], the use of tweets

to predict users getting sick[40], and detecting unknown diseases in populations[42]. In China, Weibo is a micro-blogging platform that used "sick posts" to predict disease incidence up to 14 days earlier than official public health reporting. All of these papers use post volume changes and ratios as their variable. Many papers use sentiment as a proxy for understanding disease related public opinion. Web maps have been created to show H1N1 tweet sentiment[44]. Others create custom metrics using personal negative posts to create a Degree of Concern[45], or a Measure of Concern [46]. The sentiment analysis metrics will be useful to public health officials if they can predict disease incidence in a community. The research using sentiment as a proxy for detection is not clear with one paper showing a surprising increase in sentiment polarity[22] and another showing it is inconclusive for detection[23].

Given the release of the massive geolocated twitter data set TBCOV[25] and the existing work that says Twitter can be used for disease detection. This thesis will explore the relationship between geolocated tweet sentiment and COVID-19 disease incidence at the county level. The value lies in the fact that there has been no prior investigation that compares sentiment derived from a vast collection of geolocated tweets to disease incidence data and that sentiment may offer more than the tweet count only. It is hypothesized that if there is a relationship between tweet sentiment and COVID-19 disease incidence at the local level, this can help identify patterns and trends that may be useful in predicting and tracking the spread of diseases, ultimately leading to more effective public health strategies.

Chapter 3: Data and Pre-Processing

The objectives of the thesis are to investigate the relationship between COVID-19 related tweet sentiment and official case data over space and time to determine whether tweet sentiment can be used for early detection of COVID-19. This chapter describes the data leveraged for such an analysis including the steps for pre-processing the Twitter data, the approaches used to calculate sentiment, and the collection of the COVID-19 case data.

3.1 Tweet Pre-processing

The primary source of geolocated tweets used in this work is the TBCOV dataset [25]. This dataset contains over two billion tweets collected world-wide with keywords related to COVID-19. All tweets are enriched with geolocation information, using either the location directly provided by the user or using location information from publicly available user profiles. The dataset covers a 424 day period from February 1,2020 to March 31, 2021. Since Twitter does not allow the re-distribution of Twitter data, the TBCOV dataset is provided as a "dehydrated" dataset which includes only the tweet identifiers. To obtain the full dataset, rehydration using the Twitter API is required. The code is provided to help other research to rehydrate their Twitter data in the code repository at https://github.com/jstassinos/Processing-the-TBCOV-Dataset. TBCOV is a global dataset. Therefore, all of the tweets from outside of the US are removed, reducing the number of tweets that are considered to 384,073,303. To avoid redundant information, all re-tweets from the dataset are removed. Re-tweets echo the tweet of another user and would contribute noise to the analysis. This step reduced the number of tweets to 120,678,732. Furthermore, some tweets may not be successfully rehydrated because these tweets may have already been removed - either through deletion by the user or because an account has been removed or

banned from Twitter. The tweets in this research were hydrated through April 3, 2022 to April 10, 2022. Depending on when a person hydrates the dataset the number of tweets will change. Out of the 120,678,732 unhydrated tweets, there were 77% successfully rehydrated out of tweets for a total of 93,362,576. Globally, very few tweets (0.1%) in the TBCOV dataset have explicit coordinates (latitude, longitude). Thus, for the majority of the tweets, the tweet location is inferred either based on the user location or the content of the tweet text. The authors of the TBCOV dataset report 76.1% and higher F1-scores for county-level localization using the location of the user [25]. However, for the case of locations inferred from tweet text, the F1-score was only 0.100%. For this reason, all tweets whose location is estimated using tweet text were discarded. After removing such tweets, the dataset was reduced to a total of 56,755,894 tweets. Using the set of tweets that are reliably geo-located,all tweets are grouped by day and by county. Through this process, it was observed that many counties have an insufficient number of observed tweets to constitute a representative sample. Therefore, a 7-day rolling sum of tweets is used for each county and any pairs(county, day) with fewer 30 tweets total across the seven days is discarded. The original TBCOV dataset has a gap in the tweet collection from September 15, 2020, to September 23, 2020. This period plus the 7 days following were omitted from the final aggregation.

3.2 Sentiment Measures

Sentiment is generally measured on a scale from -1 to + 1 where the signum describes the polarity (+, -) and the absolute describes the magnitude of the sentiment. To compute the sentiment of tweets for a (county, day) pair, applying four common sentiment analysis algorithms which leverage natural language processing, text analysis, and computational linguistics to systematically identify, extract, quantify, and study peoples moods, opinions, attitudes, and emotions in written text are used [47]. These are Text Blob [48], Vader Sentiment [49], Afinn [50], and an included calculation denoted as TBCOV Sentiment [24].

Text Blob [48] is a commonly used natural language processing module for Python. By

using its sentiment analysis function, it will return a named tuple with polarity and subjectivity based on the input text. TextBlob is built on top of the popular natural language processing platform Natural Language Toolkit (NLTK). The default implementation of sentiment analysis uses pattern analyzer which is built on top of pattern from NLTK. It is a Lexicon Based sentiment analyzer. TextBlob returns a score between 1 and -1 indicating both polarity (positive or negative) and strength of the sentiment [48].

Valence Aware Dictionary for Sentiment Reasoning (VADER) is the second python module that will be used to calculate sentiment on the tweets. It is a lexicon-based sentiment analyzer. This tool is also built on top of NLTK. It calculates the polarity of sentiment by matching sentiment intensity scores to words and then aggregating these scores into an overall score. This process is repeated for each tweet in the database. VADER returns a number between 1 and -1 for polarity [49].

Afinn is the third sentiment calculation python module named after its developer Finn Årup Nielsen. The developer created a dictionary of 2477 words for calculating polarity scores. Afinn is a lexicon-based sentiment analyzer. It returns a number between 5 and -5 for polarity [50].

TBCOV Sentiment, the included tweet sentiment calculation uses a multilingual transformer-based deep learning model. This returns a polarity number between 1 and -1, and a confidence score. It uses a transformer based deep learning model called XML-T, trained on millions of general-purpose tweets in 8 different languages[24]. Figure 1.1a shows the sentiment polarity from TBCOV from February 1, 2020 to March 31, 2021.

3.3 COVID-19 Case Data

The number of confirmed COVID-19 cases per county over time is collected by the Center for Disease Control (CDC) and this paper uses an aggregated version published by USA Facts[51]. After reviewing the county case data, there is additional processing required. The data is reported as a total of cases so the first step is to determine how many new

cases per day are reported Some municipalities report once a week, others will not report on weekends and this causes spikes in the new case counts[52]. To account for this a 7-day rolling average is applied. To account for population differences the daily case counts are normalized by population. When COVID-19 cases are mentioned the text is referring to a 7-day rolling average of new cases normalized by population to a per 100,000 persons value. Figure 1.1b shows the daily COVID cases from February 1, 2020 to March 31, 2021.

Chapter 4: Sentiment-Disease Case Correlation Analysis

This chapter will cover analysis methods and the data required for each method and the formulas will be formalized. Three experiments will be conducted focusing on Global correlation, temporal correlation and spatial correlation. To understand whether the correlation between COVID-19 related tweet sentiment and COVID-19 cases can be observed at the local level over space and time, the observed COVID-19-related tweets are grouped by county and calendar week. This grouping provides one set of tweets for each county r (among all counties $\mathcal{R}$)and for each week w (among all weeks $\mathcal{W}$). Deriving the sentiment of each such set provides the function $S(w, r)$ that returns the sentiment of all tweets observed in region r during week w. The following measures to quantify the correlation are defined, globally, spatially, and temporally, between these sets $S(w, r)$ and the number of COVID-19 cases $C(w, r)$ observed in the same region during the same week.

4.1 Global Analysis

To measure the correlation between COVID-19 tweet sentiment and COVID-19 cases observed over all weeks and overall regions, Pearson's correlation is used across space and time, as follows:

$$\text{corr}(\mathcal{W}, \mathcal{R}) = \text{corr}_{w \in \mathcal{W}, r \in \mathcal{R}}(S(w, r), C(w, r)) =$$

$$\frac{\sum_{w \in \mathcal{W}, r \in \mathcal{R}}(S(w, r) - \overline{S}) \cdot (C(w, r) - \overline{C})}{\sqrt{\sum_{w \in \mathcal{W}, r \in \mathcal{R}}(S(w, r) - \overline{S})^2 \cdot \sum_{w \in \mathcal{W}, r \in \mathcal{R}}(C(w, r) - \overline{C})^2}}, \tag{4.1}$$

where $\overline{S}$ denotes the average sentiment (across all weeks and counties) and $\overline{C}$ denotes the average number of cases (across all weeks and counties). Equation 4.1 provides the sample

correlation coefficient (a single scalar) that measures the correlation between COVID-19 Twitter sentiment and COVID-19 cases. Intuitively, the expectation is a negative correlation as a high number of cases should (in average across all weeks and counties) yield a low (negative) sentiment towards COVID-19. However, the experimental evaluation shows that this intuition is not confirmed by the data. That is because an aggregation across all weeks and all counties overgeneralizes any interesting spatial and temporal patterns. To find such patterns, the next proposal is to measure the correlation between COVID-19 Twitter sentiment and COVID-19 disease cases both temporally local (for a specific "frozen" week in time) and spatially local (for a specific "frozen" county).

4.2 Temporal Analysis

To understand whether and how the correlation between COVID-19 sentiment and COVID-19 cases changes over time, a correlation by week over all spatial regions is defined $\mathcal{R}$, as follows:

$$\text{corr}(w) = \text{corr}_{r \in \mathcal{R}}(S(w,r), C(w,r)) = \frac{\sum_{r \in \mathcal{R}} (S(w,r) - \overline{S(w))} \cdot (C(w,r) - \overline{C(w)})}{\sqrt{\sum_{r \in \mathcal{R}} (S(w,r) - \overline{S(w)})^2 \cdot \sum_{r \in \mathcal{R}} (C(w,r) - \overline{C(w)})^2}}, \tag{4.2}$$

where $\overline{S(w))} = \frac{\sum_{r \in \mathcal{R}} S(w,r)}{|\mathcal{R}|}$ is the average sentiment across all counties during week w and $\overline{C(w))} = \frac{\sum_{r \in \mathcal{R}} C(w,r)}{|\mathcal{R}|}$ is the average number of cases across all counties during week w.

Equation 4.2 provides the sample correlation coefficient between sentiment and cases across all states during week w. Intuitively, the expectation is that this correlation should be stationary over time, such that at any time a region having a higher number of cases should have a lower sentiment towards COVID-19. This hypothesis will be evaluated by computing corr(w) for each week and plotting the resulting time series (and corresponding p-values of the significance of the correlations) in the experiments results section.

4.3 Spatial Analysis

To understand whether the correlation between COVID-19 sentiment and COVID-19 cases is stationary over space, the correlation is computed, across all weeks $\mathcal{W}$ of the dataset for each spatial region r. This correlation is defined as:

$$\text{corr}(r) = \text{corr}_{w \in \mathcal{W}}(S(w,r), C(w,r)) = \frac{\sum_{w \in \mathcal{W}}(S(w,r) - \overline{S(r)}) \cdot (C(w,r) - \overline{C(r)})}{\sqrt{\sum_{w \in \mathcal{W}}(S(w,r) - \overline{S(r)})^2 \cdot \sum_{w \in \mathcal{W}}(C(w,r) - \overline{C(r)})^2}}, \tag{4.3}$$

where $\overline{S(r)} = \frac{\sum_{w \in \mathcal{W}} S(w,r)}{|\mathcal{W}|}$ is the average sentiment in region r over all weeks and $\overline{C(r)} = \frac{\sum_{w \in \mathcal{W}} C(w,r)}{|\mathcal{W}|}$ is the average number of cases across in region r over all weeks.

Equation 4.2 provides the sample correlation coefficient between sentiment and cases observed in region r during the entire COVID-19 pandemic. Intuitively, the expectation is that this correlation should be stationary over space, such that, for all regions, meaning that times having a higher number of cases should have a lower sentiment towards COVID-19. This hypothesis will be evaluated by computing $\text{corr}(r)$ for each region r and mapping the results across the United States in the experiments results section.

Chapter 5: Results

Recall that the thesis hypothesizes a negative association between sentiment and COVID-19 cases, where as cases go up, sentiment goes down (and vice versa). Given that this hypothesis is true, there is potential to use sentiment as a measure of early disease outbreak detection. This chapter presents the results of our correlation analysis between COVID-19-related Twitter sentiment and COVID-19 cases globally (Equation 4.1), across time (Equation 4.2), and across space (Equation 4.3).

5.1 Global Correlation

Table 5.1: Statistical values for Overall Correlation

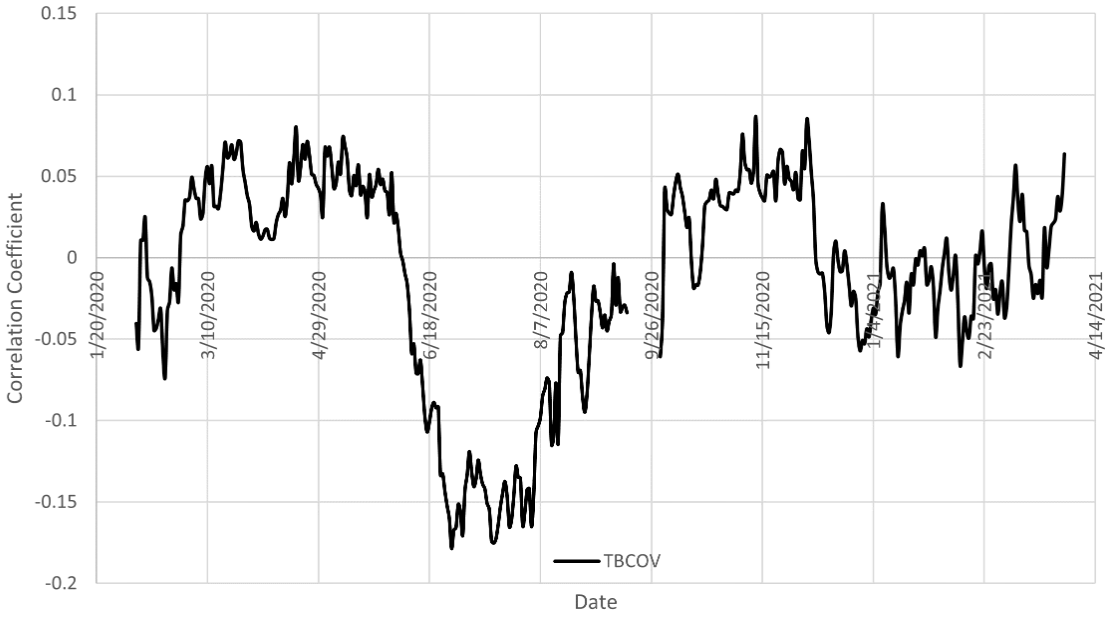
Sentiment	Correlation coefficient	T-Stat	P-Val
TBCOV	-0.05745	-44.12266	0.00000
TBLOB	-0.06347	-48.76026	0.00000
VADER	0.01663	12.75548	0.00000
AFINN	-0.02421	-18.56581	0.00000

Table 5.1 shows the global correlation between COVID-19 Twitter sentiment and COVID-19 cases using the four different sentiment measures TBCOV [25], TBLOB [48], VADER [49], and AFINN [50].

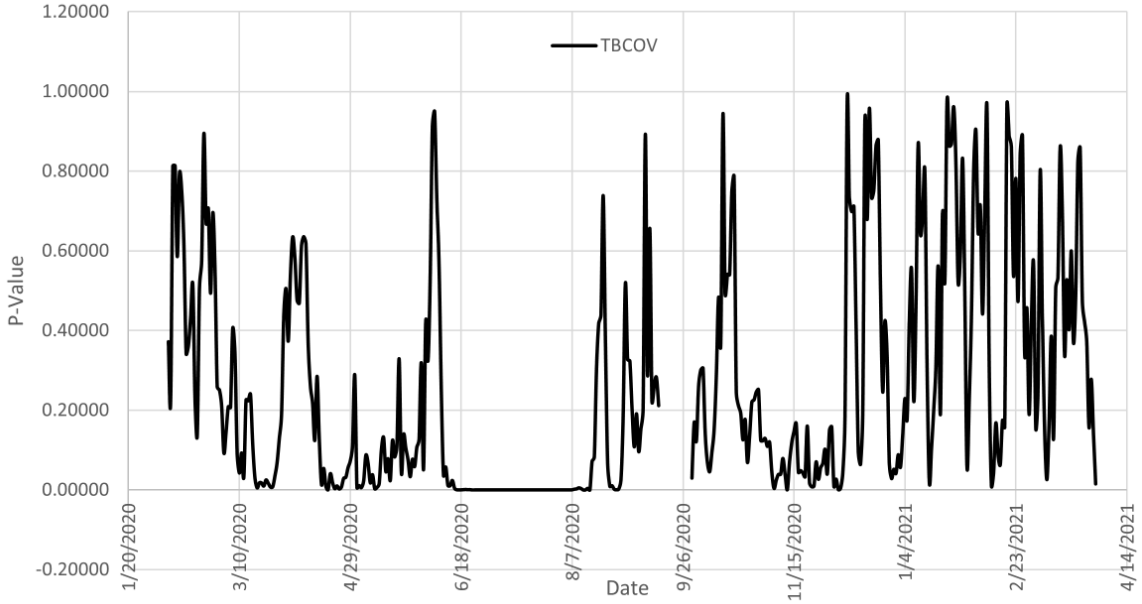
Surprisingly, the negative correlation that would be intuitively expected is observed only when using three out of the four sentiment measures. For VADER, a positive correlation

is observed. For all the others, a negative correlation is observed, but the magnitude of correlation is weak ($\leq$ 0.016). Despite the weak correlations, all these correlations are highly significantly different from zero due to the very large number of (week, county) pairs. Summarizing, Table 5.1 shows some significant correlation, but the magnitude is weak and even the direction of the correlation differs between sentiment measures. This supports the inconsistent findings from Feng and Kirkley [23] and Valdez et al. [22]. In order to examine to what degree these associations are spatially and temporally stationary and thus, may vary over time, the associations at finer spatial and temporal granularity are examined.

5.2 Correlation Over Time

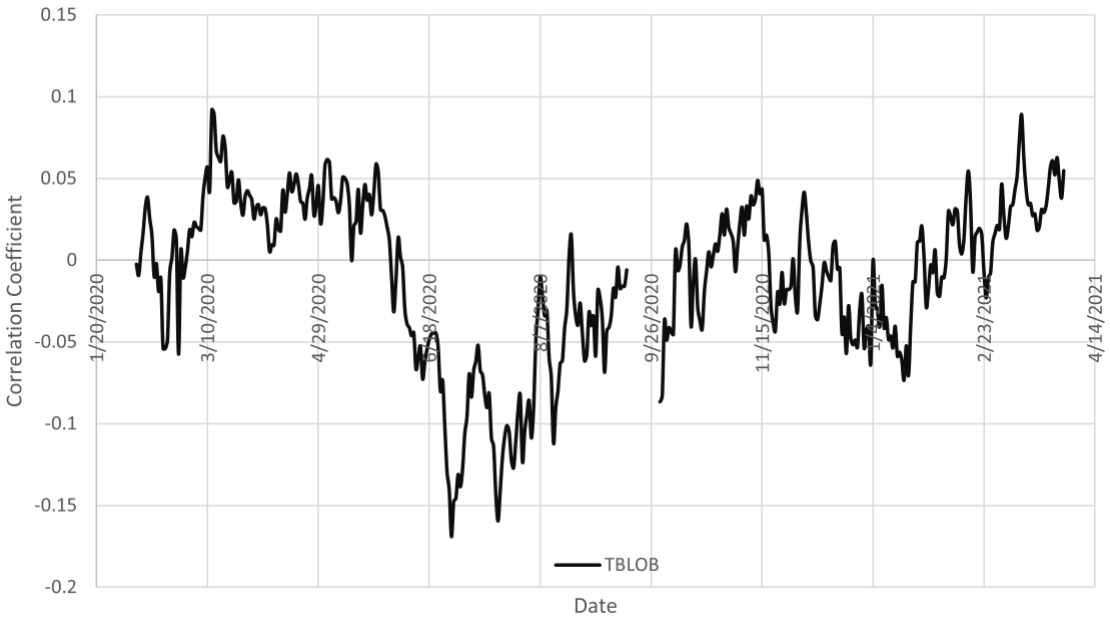


(a) correlation coefficients

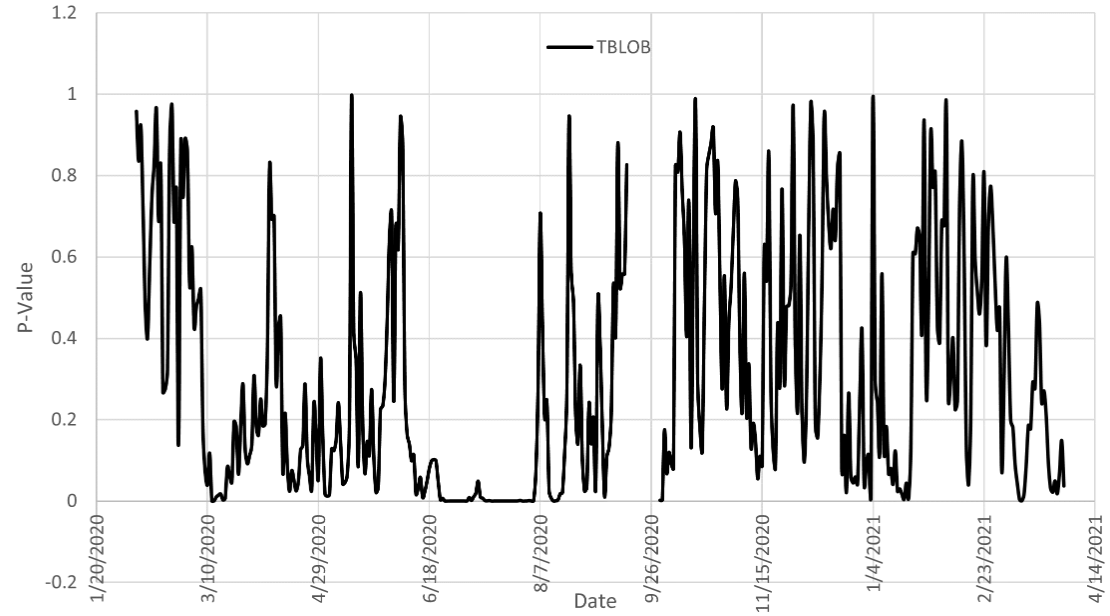


(b) p-values

Figure 5.1: (a) Daily correlation coefficient between TBCOV sentiment and the daily new cases per 100,000 in each county and (b) the corresponding p-values

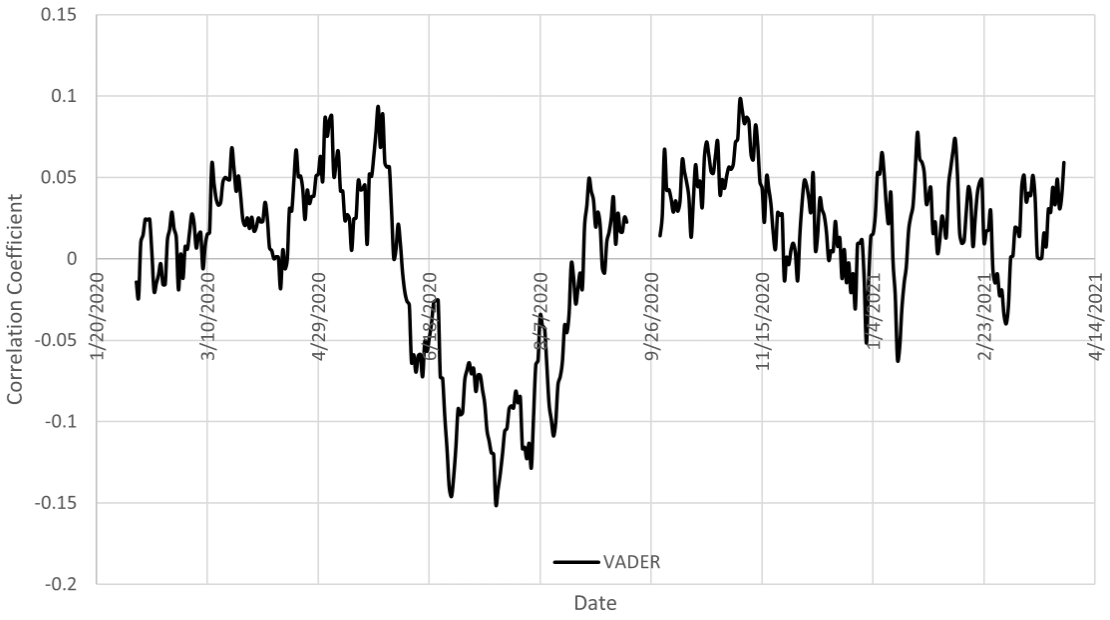


(a) correlation coefficients

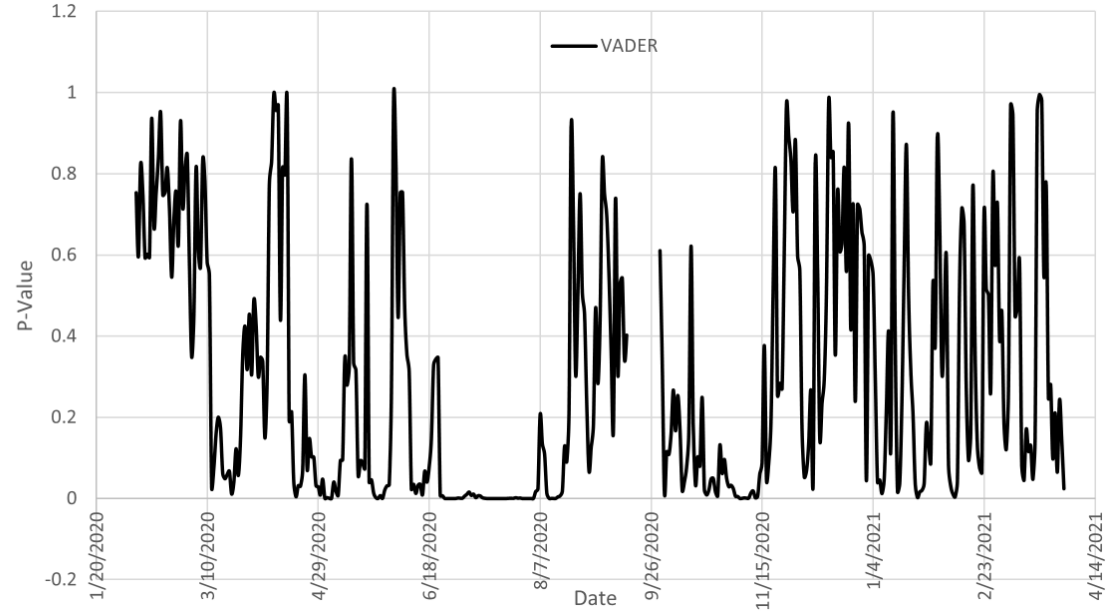


(b) p-values

Figure 5.2: (a) Daily correlation coefficient between TextBlob sentiment and the daily new cases per 100,000 in each county and (b) the corresponding p-values.

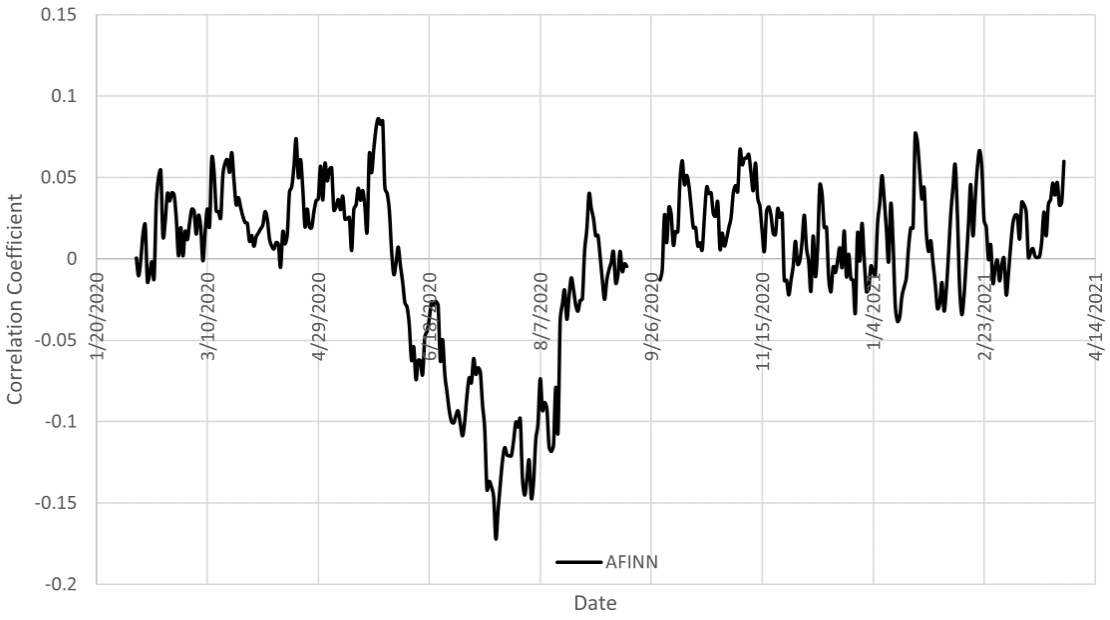


(a) correlation coefficients

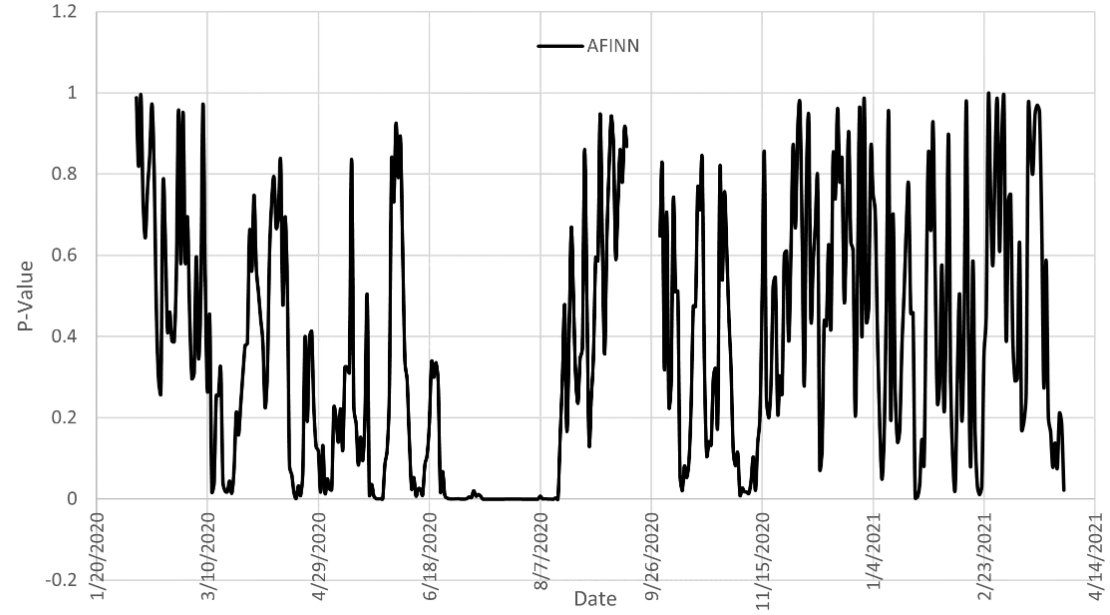


(b) p-values

Figure 5.3: (a) Daily correlation coefficient between VADER sentiment and the daily new cases per 100,000 in each county and (b) the corresponding p-values.



(a) correlation coefficients



(b) p-values

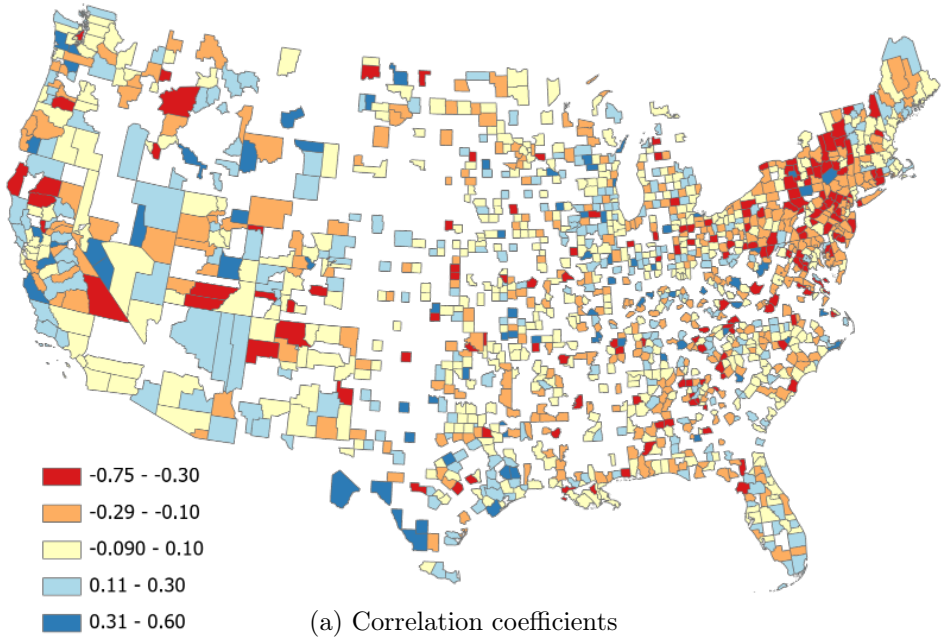
Figure 5.4: (a) Daily correlation coefficient between AFINN sentiment and the daily new cases per 100,000 in each county and (b) the corresponding p-values.

Based on Equation 4.2, Figures 5.1a, 5.2a, 5.3a, and 5.4a present the correlation coefficient between a 7-day moving average of TBCOV, TextBlob, Vader and Afinn's respective sentiment measures and a 7-day moving average of COVID-19 cases for all counties from January 20, 2020 to April 14, 2021.

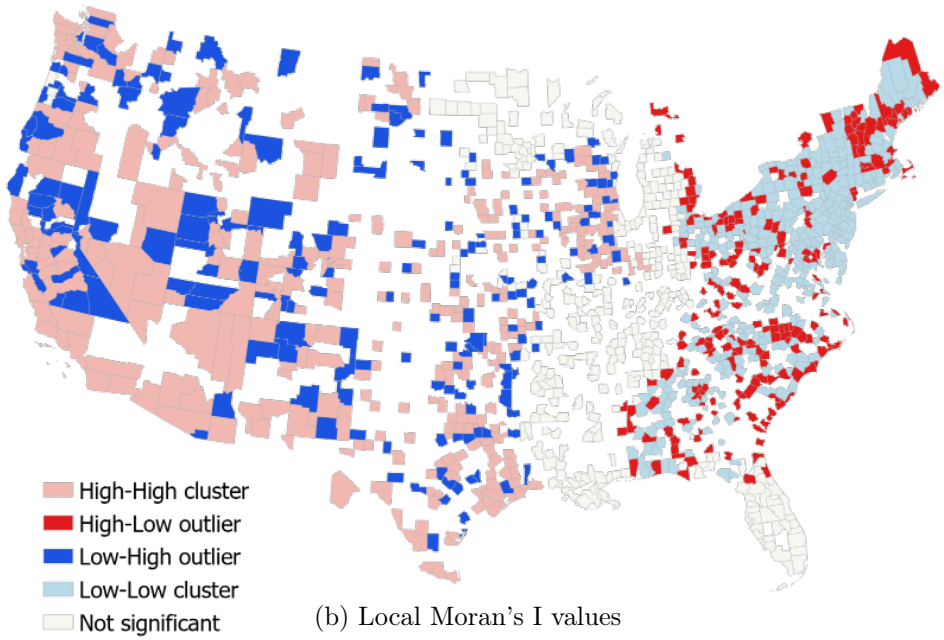
First, the hypothesis of temporal heterogeneity is supported: observing that different time intervals exhibit different magnitudes and directions of correlation between Twitter COVID-19 sentiment and COVID-19 cases. Figures 5.1b, 5.2b, 5.3b, and 5.4b provide the p-values of each correlation, thus having p-values close to zero on days where the correlation (over the last seven days) was significantly different from zero.

Looking at Figures 5.1a, 5.2a, 5.3a, and 5.4a more closely, four time intervals that exhibit significantly different correlations can be seen. The first time interval is a period of having a weakly significant (p-values < 0.05 on most days) positive correlation between tweet sentiment and COVID-19 from March 2, 2020 to June 10, 2020 where as COVID-19 cases increase, tweet sentiment increases. This unexpected positive correlation may be explained by having many counties that had zero cases at this time, but which did already exhibiting a low sentiment towards COVID-19 despite the zero cases. In late June, the correlation between sentiment and COVID-19 cases abruptly shifts to a period of highly significant (p-values approaching zero) relatively strong (0.15) negative correlation from June 12, 2020 to August 15, 2020 where as COVID-19 cases increase, tweet sentiment decreases. This short period aligns with the intuition that places with a high number of cases should have a negative sentiment towards COVID-19. Third, another shift to a period of positive correlation is observed from November 2, 2020 to December 6, 2020 similar to the first period. Finally, from December 6 onward, the correlation coefficients fluctuate around a sentiment of zero with p-values seemingly uniform in the interval $[0, 1]$ indicating no more significant correlation after this time.

5.3 Correlation Over Space

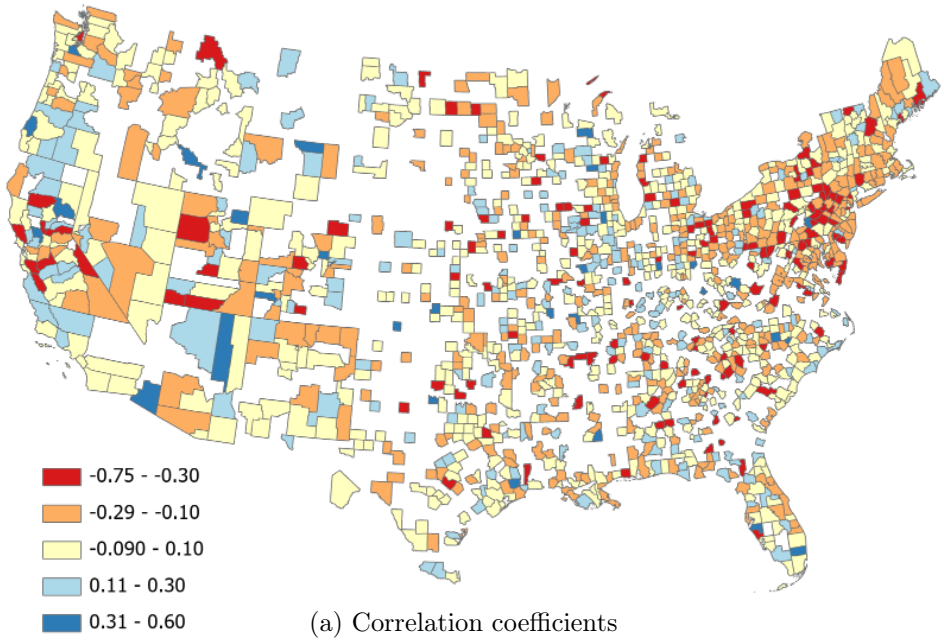


(a) Correlation coefficients

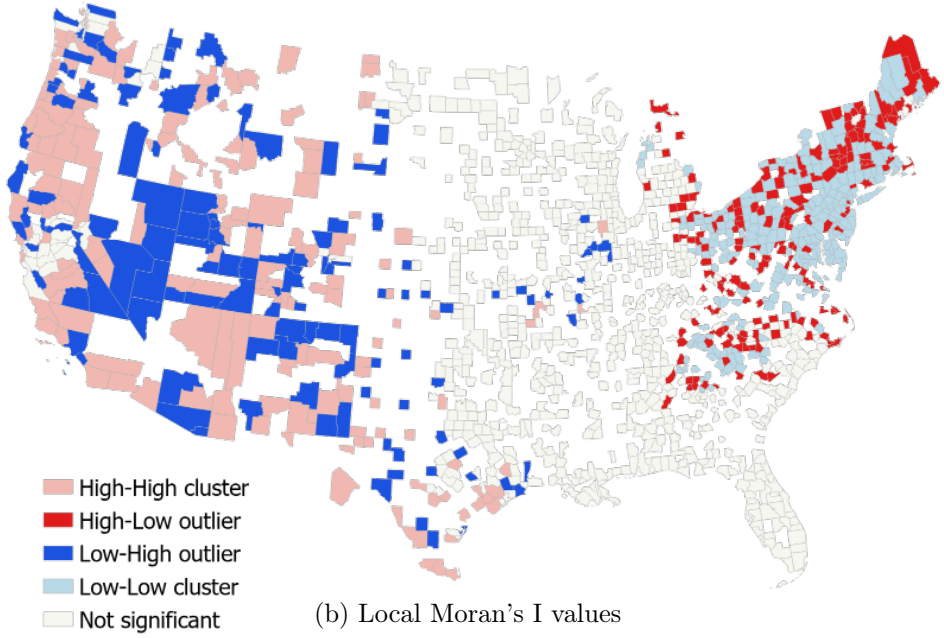


(b) Local Moran's I values

Figure 5.5: (a) Correlation coefficients between TBCOV sentiment and COVID-19 cases for all counties. (b) Anselin Local Moran's I Cluster and Outliers Analysis of TBCOV Correlation Coefficients.

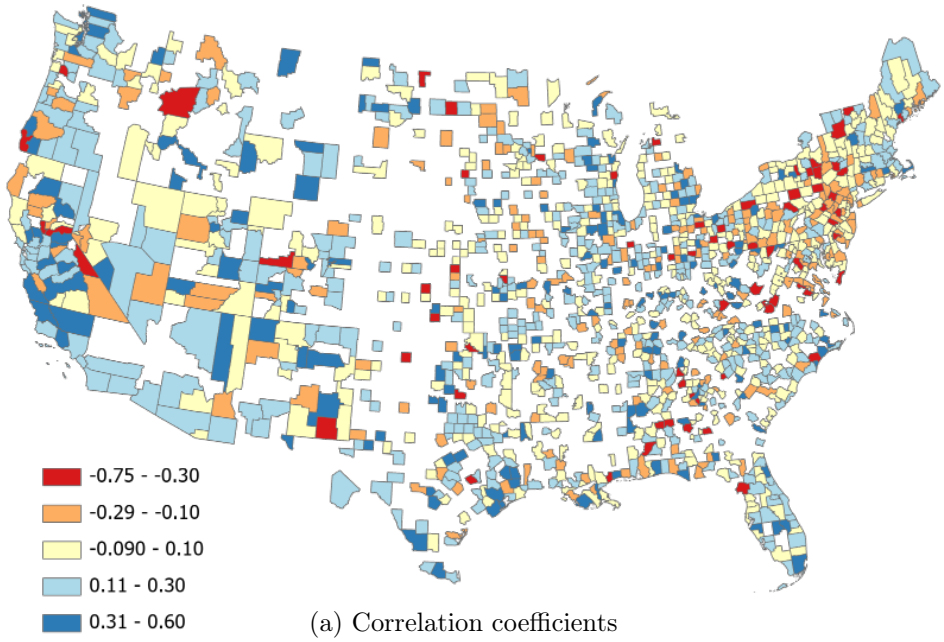


(a) Correlation coefficients

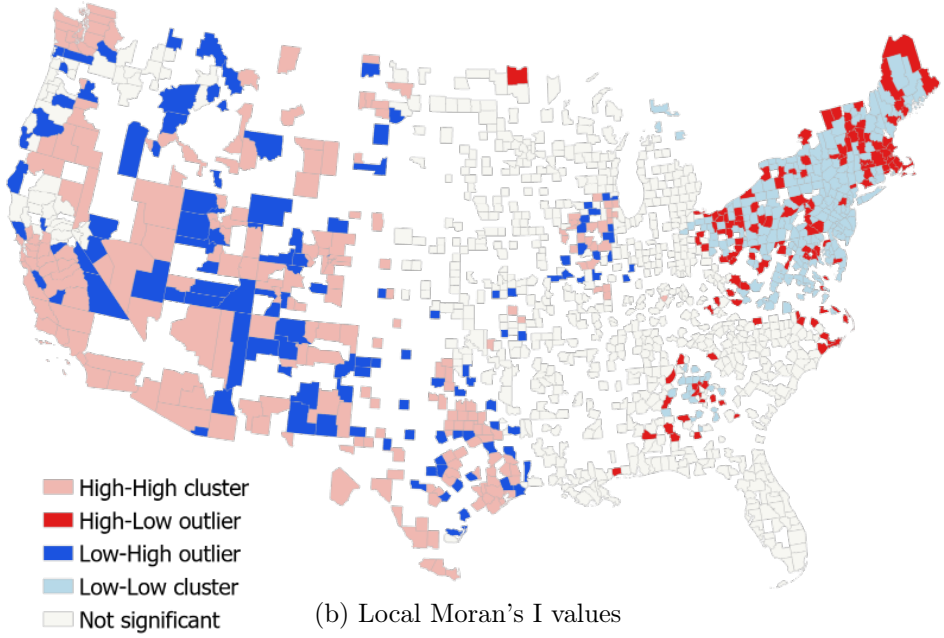


(b) Local Moran's I values

Figure 5.6: (a) Correlation coefficients between TextBlob sentiment and COVID-19 cases for all counties. (b) Anselin Local Moran's I Cluster and Outliers Analysis of TextBlob Correlation Coefficients.

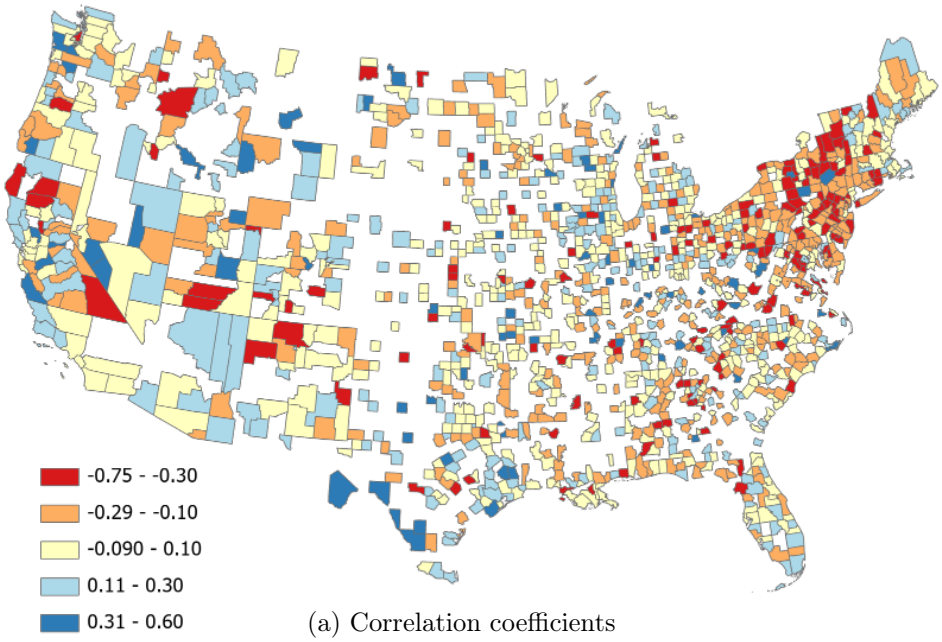


(a) Correlation coefficients

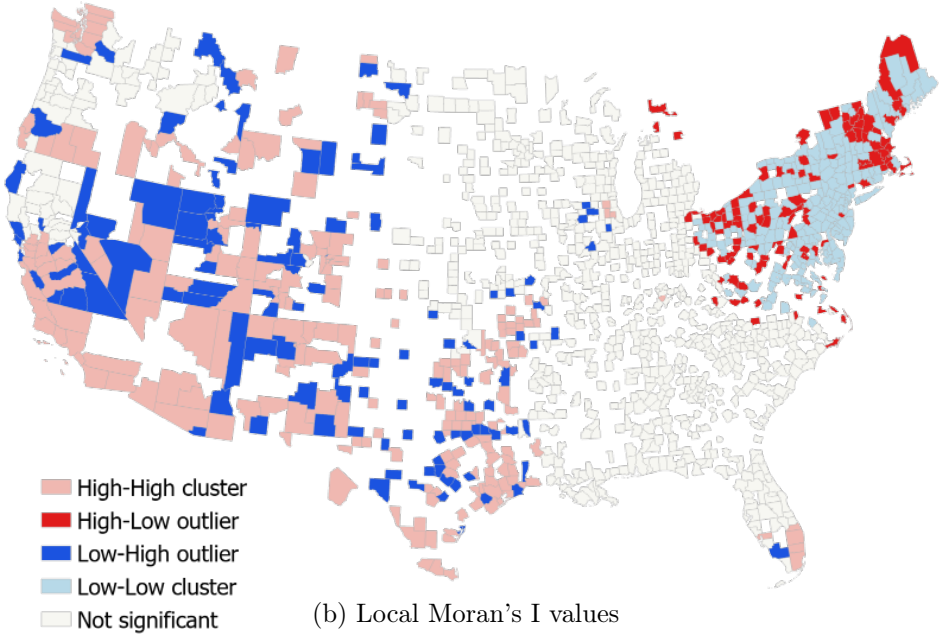


(b) Local Moran's I values

Figure 5.7: (a) Correlation coefficients between VADER sentiment and COVID-19 cases for all counties. (b) Anselin Local Moran's I Cluster and Outliers Analysis of VADER Correlation Coefficients.



(a) Correlation coefficients



(b) Local Moran's I values

Figure 5.8: (a) Correlation coefficients between AFINN sentiment and COVID-19 cases for all counties. (b) Anselin Local Moran's I Cluster and Outliers Analysis of AFINN Correlation Coefficients.

Using Equation 4.3 to obtain a correlation measure for each county, Figures 5.5a, 5.6a, 5.7a, and 5.8a presents the correlation coefficients map aggregated over all 424 days. For this experiment, any county with fewer than 200 days of data and any day with less than 30 tweets across a 7 day period was omitted. After omitting counties with insufficient data, 1,442 counties were retained out of all the 3,025 counties that have at least one tweet in the dataset. While correlations aggregated to the entire United States were rather weak, some counties did indeed exhibit strong correlations, ranging from -0.76 to 0.51. Additionally observing that these correlations exhibit spatial clustering, with many counties having either a strong negative correlation (red color in Figure 5.5a) or medium negative correlation (orange color) in the North Eastern US. Counties having a positive correlations (light and dark blue in Figure 5.5a) are less frequent, which is expected as the average correlation for TBCOV is -0.055 as shown in Table 5.1, but appear more frequently in the West and Midwest. As these spatial trends may not be immediately evident from Figure 5.5a, applying Anselin's test for local spatial autocorrelation to the correlations values obtaining the map of local clusters as shown in Figure 5.5b, 5.6b, 5.7b, and 5.8b. As expected, large cluster of low values is observed (this indicating negative correlation values) to the Northeast (although having many outliers of relatively high values within this low cluster). A large cluster of high values is observed (indicating positive correlation) in the West and Midwest, but with parts of California being excluded from this cluster as not significant.

Chapter 6: Discussion and Conclusions

The objective of this thesis was to determine whether infectious disease-related tweet sentiment (for which billions of tweets have been made available for analysis) at a given time and location exhibits an inverse correlation to the number of observed disease cases at the same time and location. If such a correlation could be shown, it may be possible to use local sentiment as an early indicator of an outbreak that would precede reported cases and thus, may give local health departments a valuable head start to save lives and curb disease spread.

Globally, across all counties of the United States and across the entire one-year study period, report a negative correlation coefficient. The correlation is very weak, and even shows different directions (positive and negative) for different sentiment measures. When conducting Pearson's correlation analysis, there is a maximum number of data points that can be used before the statistical significance of the correlation coefficient, as indicated by the p-value, becomes unreliable or meaningless. Thus, the analysis is extended to spatial and temporal dimensions to investigate the spatial and temporal stationarity of such a trend and whether variations in the magnitude and direction of the correlation could be found at certain times or in certain locations.

The temporal analysis showed that the inconclusive global correlation exhibits significant temporal patterns, including an interval of a strong and significant negative correlation in the Summer of 2020. Thus, during this time, Twitter sentiment could have indeed been used as a significant indicator of the severity of COVID-19 cases locally. There are also periods of positive correlation (which contradicts the intuition that a high number of cases should correspond with low sentiment) as well as periods of insignificant correlations. This result is quite interesting: It shows that if it is possible to understand the type of correlation (positive,

negative, strong, weak, spatial, temporal) during a new infectious disease outbreak, Twitter sentiment could indeed be leveraged as a local indicator.

A spatial analysis was performed to understand whether different counties across the United States may also exhibit non-stationarity. It was observed that areas in the Northeast of the U.S. exhibit a stronger negative correlation while West exhibits are more positive correlation. This result is also interesting, as it shows some regions of the U.S. may allow more accurate forecasting of infectious disease cases based on Twitter sentiment than others. Further research is needed to understand both temporal and spatial processes that cause the correlations between Twitter sentiment and infectious diseases to shift across space and time. In understanding these processes, it may be possible to identify when and where Twitter sentiment could be used as a signal for early disease outbreak warning.

There are some limitations to this thesis, specifically emanating from the spatial data and the availability of tweet data. First, the original dataset starts with approximately 384 million tweets in the Unites States but not all of these tweets are able to be geolocated in a county, additionally some counties do not have enough tweets to consider in the analysis. This leaves gaps in the analysis with sparsely populated counties underrepresented. A similar situation can occur where one user is doing the majority of the tweeting for a county causing their individual sentiments to represent all the users in that county. Furthermore, it is more likely that a user location is identified if they are a more prolific user of the platform, increasing the likelihood that a few users are over represented in a counties data[53,54].Second, the modifiable areal unit problem becomes an issue when using counties as the container for tweets and cases which are point features. The distribution of the population within a county is not uniform but when attaching tweets and cases to a county it is represented as uniform. Third, sentiment measures can also be a limitation, in text with complex linguistics or common phrasing of a topic favoring positive or negative polarity. Finally, correlation values on their own carry no meaning as they are unitless, future work would need to determine the level of correlation that has meaning.

Chapter 7: Contributions

This thesis addresses a gap in the existing knowledge bases of Geography, Data Science, and Public Health. Before this thesis there was no research to our knowledge that directly compared local twitter sentiment to local disease incidence with the purpose of local disease prediction.

Health geography uses spatial information and analysis techniques to study health and disease. This thesis specifically focuses on the spatial distribution of the COVID-19 cases and the local tweet sentiment and the relationship between the two. While the results are not without limitations they still provide evidence for sentiment representing real-world conditions at certain times and places. The results provide justification for monitoring DOC and MOC metrics elevating their significance beyond an academic exercise. Furthermore, the spatial non-stationary nature of this relationship as seen in Figures 5.5, 5.6, 5.7, and 5.8 requires further investigation and in future research identifying what causes the spatial variations in correlation will improve the model. Furthermore, identifying what topic specifically caused the sentiment in the corpus using topic sentiment analysis can further improve the results [55, 56].

The field of sentiment analysis is a subfield of Natural Language Processing (NLP), which investigates the interactions between computers and human language, with a specific focus on the analysis of human emotions and sentiments. In this thesis four different python libraries are utilized[25, 48–50] to process the geolocated tweet corpus with each of them showing a similar correlations with COVID-19 as seen in Figures 5.1, 5.2, 5.3, and 5.4. This thesis provides evidence for the partial efficacy of applying sentiment analysis techniques on a large corpus of tweets, as they can reflect real world conditions at certain times and places. With the continuous advancements in NLP techniques, sentiment analysis will likely improve in accuracy and effectiveness.

Disease detection in public health is an essential part of mitigating the spread of infections in a population and accurate local disease incidence information will enable public health decision makers to predict outbreaks and deploy mitigation strategies more effectively. An improvement in understanding how tweet sentiment correlates with disease incidence can improve disease modeling. This thesis clearly show that at some times and in some areas twitter sentiment can be used for disease prediction, and therefore this thesis is contributing to improving the use of sentiment analysis for disease detection.

Bibliography

[1] D. Bertsimas, L. Boussioux, R. Cory-Wright, A. Delarue, V. Digalakis, A. Jacquillat, D. L. Kitane, G. Lukin, M. Li, L. Mingardi *et al.*, "From predictions to prescriptions: A data-driven response to covid-19," *Health care management science*, vol. 24, no. 2, pp. 253–272, 2021.

[2] J. Elarde, J.-S. Kim, H. Kavak, A. Züfle, and T. Anderson, "Change of human mobility during covid-19: A united states case study," *PloS one*, vol. 16, no. 11, p. e0259031, 2021.

[3] CDC, "Cases, Data, and Surveillance," Feb. 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/about-us-cases-deaths.html

[4] G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. Davey Smith *et al.*, "Collider bias undermines our understanding of covid-19 disease risk and severity," *Nature communications*, vol. 11, no. 1, p. 5749, 2020.

[5] B. Rader, C. M. Astley, K. T. L. Sy, K. Sewalk, Y. Hswen, J. S. Brownstein, and M. U. Kraemer, "Geographic access to united states sars-cov-2 testing sites highlights healthcare disparities and may bias transmission estimates," *Journal of travel medicine*, vol. 27, no. 7, p. taaa076, 2020.

[6] R. Jin, "The lag between daily reported covid-19 cases and deaths and its relationship to age," *Journal of Public Health Research*, vol. 10, no. 3, pp. jphr–2021, 2021.

[7] A. C. Nagel, M.-H. Tsou, B. H. Spitzberg, L. An, J. M. Gawron, D. K. Gupta, J.-A. Yang, S. Han, K. M. Peddecord, S. Lindsay *et al.*, "The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets," *Journal of medical Internet research*, vol. 15, no. 10, p. e2705, 2013.

[8] A. E. Aiello, A. Renson, and P. Zivich, "Social media-and internet-based disease surveillance for public health," *Annual review of public health*, vol. 41, p. 101, 2020.

[9] E. B. Kpozehouen, X. Chen, M. Zhu, and C. R. Macintyre, "Using open-source intelligence to detect early signals of covid-19 in china: descriptive study," *JMIR Public Health and Surveillance*, vol. 6, no. 3, p. e18939, 2020.

[10] S. Nair, A. Moa, and R. Macintyre, "Investigation of early epidemiological signals of covid-19 in india using outbreak surveillance data," *Global Biosecurity*, vol. 2, no. 1, 2020.

[11] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection—harnessing the web for public health surveillance," *The New England journal of medicine*, vol. 360, no. 21, p. 2153, 2009.

[12] R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara, J. S. Brownstein *et al.*, "A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives," *Journal of medical Internet research*, vol. 16, no. 10, p. e3416, 2014.

[13] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic," *PloS one*, vol. 8, no. 12, p. e83672, 2013.

[14] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1474–1477.

[15] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*. IEEE, 2011, pp. 702–707.

[16] M. Szomszor, P. Kostkova, and E. d. Quincey, "# swineflu: Twitter predicts swine flu outbreak in 2009," in *International conference on electronic healthcare*. Springer, 2010, pp. 18–26.

[17] J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proceedings of the 3rd international web science conference*, 2011, pp. 1–8.

[18] V. Carchiolo, A. Longheu, and M. Malgeri, "Using twitter data and sentiment analysis to study diseases dynamics," in *International conference on information technology in bio-and medical informatics*. Springer, 2015, pp. 16–24.

[19] K. Byrd, A. Mansurov, and O. Baysal, "Mining twitter data for influenza detection and surveillance," in *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, 2016, pp. 43–49.

[20] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter," in *Proceedings of the 2011 Conference on empirical methods in natural language processing*, 2011, pp. 1568–1576.

[21] S. Doan, L. Ohno-Machado, and N. Collier, "Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses," in *2012 iEEE second international conference on healthcare informatics, imaging and systems biology*. IEEE, 2012, pp. 62–71.

[22] D. Valdez, M. Ten Thij, K. Bathina, L. A. Rutter, J. Bollen *et al.*, "Social media insights into us mental health during the covid-19 pandemic: Longitudinal analysis of twitter data," *Journal of medical Internet research*, vol. 22, no. 12, p. e21418, 2020.

[23] S. Feng and A. Kirkley, "Integrating online and offline data for crisis management: Online geolocalized emotion, policy response, and local mobility during the covid crisis," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[24] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3404820.3404823

[25] M. Imran, U. Qazi, and F. Ofli, "TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels," *Data*, vol. 7, no. 1, p. 8, Jan. 2022, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2306-5729/7/1/8

[26] D. Koutsoyiannis and A. Montanari, "Negligent killing of scientific concepts: the stationarity case," *Hydrological Sciences Journal*, vol. 60, no. 7-8, pp. 1174–1183, 2015.

[27] S. M. Teutsch, R. E. Churchill *et al.*, *Principles and practice of public health surveillance*. Oxford University Press, USA, 2000.

[28] J. Snow, *On the mode of communication of cholera*. John Churchill, 1849.

[29] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[30] C. D. Corley, A. R. Mikler, K. P. Singh, and D. J. Cook, "Monitoring influenza trends through mining social media." *BIOCOMP*, vol. 2009, pp. 340–6, 2009.

[31] N. Beauchamp, ""this candle has no smell": Detecting the effect of covid anosmia on amazon reviews using bayesian vector autoregression," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1363–1367.

[32] M. Mokbel, M. Sakr, L. Xiong, A. Züfle, J. Almeida, T. Anderson, W. Aref, G. Andrienko, N. Andrienko, Y. Cao *et al.*, "Mobility data science (dagstuhl seminar 22021)," in *Dagstuhl reports*, vol. 12, no. 1. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[33] M. T. Le, D. Attaway, T. Anderson, H. Kavak, A. Roess, and A. Züfle, "Phyloview: A system to visualize the ecology of infectious diseases using phylogenetic data," in *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2022, pp. 222–229.

[34] Z. Andreas, S. Gao, and T. Anderson, "Introduction to the special issue on understanding the spread of covid-19, part 2," pp. 1–5, 2022.

[35] A. Züfle, T. Anderson, and S. Gao, "Introduction to the special issue on understanding the spread of covid-19, part 1," pp. 1–5, 2022.

[36] M. F. Mokbel, L. Xiong, and D. Zeinalipour-Yazti, "Introduction to the special issue on contact tracing," pp. 1–2, 2022.

[37] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle, "Data-driven mobility models for covid-19 simulation," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, 2020, pp. 29–38.

[38] J.-S. Kim, H. Jin, and A. Züfle, "Expert-in-the-loop prescriptive analytics using mobility intervention for epidemics," in *Proceedings of the First Workshop on Prescriptive Analytics for the Physical World (PAPW 2020)(KDD2020)*, 2020.

[39] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, pp. 211–221, 2007.

[40] A. Sadilek, H. Kautz, and V. Silenzio, "Predicting disease transmission from geo-tagged micro-blog data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 136–142.

[41] M. Paul, "Dredze m.(2011b). you are what you tweet: Analyzing twitter for public health," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2013, pp. 265–272.

[42] S. Lim, C. S. Tucker, and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *Journal of biomedical informatics*, vol. 66, pp. 82–94, 2017.

[43] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, W. Liao *et al.*, "Using reports of symptoms and diagnoses on social media to predict covid-19 case counts in mainland china: Observational infoveillance study," *Journal of medical Internet research*, vol. 22, no. 5, p. e19421, 2020.

[44] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[45] P. N. Behera, S. Eluri *et al.*, "Analysis of public health concerns using two-step sentiment classification," *Int. J. Eng. Res. Technol*, vol. 4, pp. 606–610, 2015.

[46] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," *Social Network Analysis and Mining*, vol. 5, pp. 1–25, 2015.

[47] B. Liu *et al.*, "Sentiment analysis and subjectivity." *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[48] S. Loria, "TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation," 2020. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[49] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[50] F. Å. Nielsen, "afinn," Apr. 2022, original-date: 2015-07-06T08:34:13Z. [Online]. Available: https://github.com/fnielsen/afinn

[51] "US COVID-19 cases and deaths by state," Oct. 2021. [Online]. Available: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

[52] CDC, "COVID Data Tracker," Mar. 2020. [Online]. Available: https://covid.cdc.gov/covid-data-tracker

[53] E. Seglem, A. Züfle, J. Stutzki, F. Borutta, E. Faerman, and M. Schubert, "On privacy in spatio-temporal data: User identification using microblog data," in *Advances in Spatial and Temporal Databases: 15th International Symposium, SSTD 2017, Arlington, VA, USA, August 21–23, 2017, Proceedings 15*. Springer, 2017, pp. 43–61.

[54] J. D. Park, E. Seglem, E. Lin, and A. Züfle, "Protecting user privacy: Obfuscating discriminative spatio-temporal footprints," in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Recommendations for Location-based Services and Social Networks*, 2017, pp. 1–4.

[55] G. Prathap, H. Kavak, E. Kaya, L. Palmieri, S. Karahan, and A. Korb, "Anti-american stance in turkey: A twitter case study," in *International Conference on Cyber Warfare and Security*, vol. 18, no. 1, 2023, pp. 309–317.

[56] R. Korb, S. Karahan, G. Prathap, E. Kaya, L. Palmieri, H. Kavak *et al.*, "S-400s, disinformation, and anti-american sentiment in turkey," 2023.